

Basic Business Statistics

13th
Edition

Concepts and Applications



Berenson

Levine

Szabat

A ROADMAP FOR SELECTING A STATISTICAL METHOD

Data Analysis Task	For Numerical Variables	For Categorical Variables
Describing a group or several groups	<p>Ordered array, stem-and-leaf display, frequency distribution, relative frequency distribution, percentage distribution, cumulative percentage distribution, histogram, polygon, cumulative percentage polygon, bullet maps, sparklines, gauges, treemaps (Sections 2.2, 2.4, 17.1)</p> <p>Mean, median, mode, geometric mean, quartiles, range, interquartile range, standard deviation, variance, coefficient of variation, skewness, kurtosis, boxplot, normal probability plot (Sections 3.1, 3.2, 3.3, 6.3)</p> <p>Index numbers (online Section 16.8)</p> <p>Gauges, bullet graphs, and treemaps (Section 17.1)</p>	Summary table, bar chart, pie chart, Pareto chart (Sections 2.1 and 2.3)
Inference about one group	<p>Confidence interval estimate of the mean (Sections 8.1 and 8.2)</p> <p><i>t</i> test for the mean (Section 9.2)</p> <p>Chi-square test for a variance or standard deviation (online Section 12.7)</p>	<p>Confidence interval estimate of the proportion (Section 8.3)</p> <p>Z test for the proportion (Section 9.4)</p>
Comparing two groups	<p>Tests for the difference in the means of two independent populations (Section 10.1)</p> <p>Wilcoxon rank sum test (Section 12.4)</p> <p>Paired <i>t</i> test (Section 10.2)</p> <p><i>F</i> test for the difference between two variances (Section 10.4)</p> <p>Wilcoxon signed ranks test (online Section 12.8)</p>	<p>Z test for the difference between two proportions (Section 10.3)</p> <p>Chi-square test for the difference between two proportions (Section 12.1)</p> <p>McNemar test for two related samples (online Section 12.6)</p>
Comparing more than two groups	<p>One-way analysis of variance for comparing several means (Section 11.1)</p> <p>Kruskal-Wallis test (Section 12.5)</p> <p>Randomized block design (Section 11.2)</p> <p>Two-way analysis of variance (Section 11.3)</p> <p>Friedman rank test (online Section 12.9)</p>	Chi-square test for differences among more than two proportions (Section 12.2)
Analyzing the relationship between two variables	<p>Scatter plot, time series plot (Section 2.5)</p> <p>Covariance, coefficient of correlation (Section 3.5)</p> <p>Simple linear regression (Chapter 13)</p> <p><i>t</i> test of correlation (Section 13.7)</p> <p>Time-series forecasting (Chapter 16)</p> <p>Sparklines (Section 17.1)</p>	<p>Contingency table, side-by-side bar chart, PivotTables (Sections 2.1, 2.3, 2.6)</p> <p>Chi-square test of independence (Section 12.3)</p>
Analyzing the relationship between two or more variables	<p>Multiple regression (Chapters 14 and 15)</p> <p>Regression trees (Section 17.3)</p> <p>Neural nets (Section 17.4)</p> <p>Cluster analysis (Section 17.5)</p> <p>Multidimensional scaling (Section 17.6)</p>	<p>Multidimensional contingency tables (Section 2.7)</p> <p>Drilldown and slicers (Section 17.1)</p> <p>Logistic regression (Section 14.7)</p> <p>Classification trees (Section 17.3)</p> <p>Neural nets (Section 17.4)</p>

This page intentionally left blank

Basic Business Statistics

Concepts and Applications

THIRTEENTH EDITION

Mark L. Berenson

Department of Information and Operations Management

School of Business, Montclair State University

David M. Levine

Department of Statistics and Computer Information Systems

Zicklin School of Business, Baruch College, City University of New York

Kathryn A. Szabat

Department of Business Systems and Analytics

School of Business, La Salle University

PEARSON

Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editor in Chief: Deirdre Lynch
Acquisitions Editor: Marianne Stepanian
Project Editor: Dana Bettez
Assistant Editor: Sonia Ashraf
Senior Managing Editor: Karen Wernholm
Senior Production Supervisor: Kathleen A. Manley
Digital Assets Manager: Marianne Groth
Manager, Multimedia Production: Christine Stavrou
Software Development: John Flanagan, MathXL; Marty Wright, TestGen
Senior Marketing Manager: Erin Lane

Marketing Assistant: Kathleen DeChavez
Senior Author Support/Technology Specialist: Joe Vetere
Rights and Permissions Advisor: Cathy Pare
Image Manager: Rachel Youdelman
Procurement Specialist: Debbie Rossi
Art Direction and Cover Design: Barbara Atkinson
Text Design, Production Coordination, Composition, and Illustrations: PreMediaGlobal
Cover photo several-businesspeople-walking-in-the-corridor: Pressmaster/Shutterstock

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Pearson was aware of a trademark claim, the designations have been printed in initial caps or all caps.

MICROSOFT® AND WINDOWS® ARE REGISTERED TRADEMARKS OF THE MICROSOFT CORPORATION IN THE U.S.A. AND OTHER COUNTRIES. THIS BOOK IS NOT SPONSORED OR ENDORSED BY OR AFFILIATED WITH THE MICROSOFT CORPORATION. ILLUSTRATIONS OF MICROSOFT EXCEL IN THIS BOOK HAVE BEEN TAKEN FROM MICROSOFT EXCEL 2013, UNLESS OTHERWISE INDICATED.

MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS MAKE NO REPRESENTATIONS ABOUT THE SUITABILITY OF THE INFORMATION CONTAINED IN THE DOCUMENTS AND RELATED GRAPHICS PUBLISHED AS PART OF THE SERVICES FOR ANY PURPOSE. ALL SUCH DOCUMENTS AND RELATED GRAPHICS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND. MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS HEREBY DISCLAIM ALL WARRANTIES AND CONDITIONS WITH REGARD TO THIS INFORMATION, INCLUDING ALL WARRANTIES AND CONDITIONS OF MERCHANTABILITY, WHETHER EXPRESS, IMPLIED OR STATUTORY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT. IN NO EVENT SHALL MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF INFORMATION AVAILABLE FROM THE SERVICES.

THE DOCUMENTS AND RELATED GRAPHICS CONTAINED HEREIN COULD INCLUDE TECHNICAL INACCURACIES OR TYPOGRAPHICAL ERRORS. CHANGES ARE PERIODICALLY ADDED TO THE INFORMATION HEREIN. MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS MAY MAKE IMPROVEMENTS AND/OR CHANGES IN THE PRODUCT(S) AND/OR THE PROGRAM(S) DESCRIBED HEREIN AT ANY TIME. PARTIAL SCREEN SHOTS MAY BE VIEWED IN FULL WITHIN THE SOFTWARE VERSION SPECIFIED.

Minitab © 2013. Portions of information contained in this publication/book are printed with permission of Minitab Inc. All such material remains the exclusive property and copyright of Minitab Inc. All rights reserved.

The contents, descriptions, and characters of WaldoLands and Waldowood are Copyright © 2014, 2011 Waldowood Productions, and used with permission.

Library of Congress Cataloging-in-Publication Data

Berenson, Mark L.
Basic Business Statistics / Mark L. Berenson, David M. Levine, Kathryn A. Szabat.—13th ed.

p. cm.

ISBN 978-0-321-87002-5

1. Commercial statistics. 2. Industrial management—Statistical methods. I. Levine, David M. II. Szabat, Kathryn, A. III. Title.
HF1017.S74 2013
519.5024'65—dc23

Copyright © 2015, 2012, 2009 Pearson Education, Inc. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America. For information on obtaining permission for use of material in this work, please submit a written request to Pearson Education, Inc., Rights and Contracts Department, 501 Boylston Street, Suite 900, Boston, MA 02116, fax your request to 617-671-3447, or e-mail at <http://www.pearsoned.com/legal/permissions.htm>.

1 2 3 4 5 6 7 8 9 10—CRK—17 16 15 14 13

PEARSON

www.pearsonhighered.com

ISBN-10: 0-321-87002-6
ISBN-13: 978-0-321-87002-5

*To our spouses and children,
Rhoda, Marilyn, Kathy, Lori, and Sharyn*

*and to our parents, in loving memory,
Nat, Ethel, Lee, Reuben, Mary, and William*

About the Authors



The authors of this book: Kathryn Szabat, David Levine, and Mark Berenson at a Decision Sciences Institute meeting.

Mark L. Berenson is Professor of Management and Information Systems at Montclair State University (Montclair, New Jersey) and also Professor Emeritus of Statistics and Computer Information Systems at Bernard M. Baruch College (City University of New York). He currently teaches graduate and undergraduate courses in statistics and in operations management in the School of Business and an undergraduate course in international justice and human rights that he co-developed in the College of Humanities and Social Sciences.

Berenson received a B.A. in economic statistics and an M.B.A. in business statistics from City College of New York and a Ph.D. in business from the City University of New York. Berenson's research has been published in *Decision Sciences Journal of Innovative Education*, *Review of Business Research*, *The American Statistician*, *Communications in Statistics*, *Psychometrika*, *Educational and Psychological Measurement*, *Journal of Management Sciences and Applied Cybernetics*, *Research Quarterly*, *Stats Magazine*, *The New York Statistician*, *Journal of Health Administration Education*, *Journal of Behavioral Medicine*, and *Journal of Surgical Oncology*. His invited articles have appeared in *The Encyclopedia of Measurement & Statistics* and *Encyclopedia of Statistical Sciences*. He is co-author of 11 statistics texts published by Prentice Hall, including *Statistics for Managers Using Microsoft Excel*, *Basic Business Statistics: Concepts and Applications*, and *Business Statistics: A First Course*.

Over the years, Berenson has received several awards for teaching and for innovative contributions to statistics education. In 2005, he was the first recipient of the Catherine A. Becker Service for Educational Excellence Award at Montclair State University and, in 2012, he was the recipient of the Khubani/Telebrands Faculty Research Fellowship in the School of Business.

David M. Levine is Professor Emeritus of Statistics and Computer Information Systems at Baruch College (City University of New York). He received B.B.A. and M.B.A. degrees in statistics from City College of New York and a Ph.D. from New York University in industrial engineering and operations research. He is nationally recognized as a leading innovator in statistics education and is the co-author of 14 books, including such best-selling statistics textbooks as *Statistics for Managers Using Microsoft Excel*, *Basic Business Statistics: Concepts and Applications*, *Business Statistics: A First Course*, and *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*.

He also is the co-author of *Even You Can Learn Statistics: A Guide for Everyone Who Has Ever Been Afraid of Statistics*, currently in its second edition, *Six Sigma for Green Belts and Champions* and *Design for Six Sigma for Green Belts and Champions*, and the author of *Statistics for Six Sigma Green Belts*, all published by FT Press, a Pearson imprint, and *Quality Management*, third edition, McGraw-Hill/Irwin. He is also the author of *Video Review of Statistics* and *Video Review of Probability*, both published by Video Aided Instruction, and the statistics module of the MBA primer published by Cengage Learning. He has published articles in various journals, including *Psychometrika*, *The American Statistician*, *Communications in Statistics*, *Decision Sciences Journal of Innovative Education*, *Multivariate Behavioral Research*, *Journal of Systems Management*, *Quality Progress*, and *The American Anthropologist*, and he has given numerous talks at the Decision Sciences Institute (DSI), American Statistical Association (ASA), and Making Statistics More Effective in Schools and Business (MSMESB) conferences. Levine has also received several awards for outstanding teaching and curriculum development from Baruch College.

Kathryn A. Szabat is Associate Professor and Chair of Business Systems and Analytics at LaSalle University. She teaches undergraduate and graduate courses in business statistics and operations management.

Szabat's research has been published in *International Journal of Applied Decision Sciences*, *Accounting Education*, *Journal of Applied Business and Economics*, *Journal of Healthcare Management*, and *Journal of Management Studies*. Scholarly chapters have appeared in *Managing Adaptability, Intervention, and People in Enterprise Information Systems*; *Managing, Trade, Economics and International Business*; *Encyclopedia of Statistics in Behavioral Science*; and *Statistical Methods in Longitudinal Research*.

Szabat has provided statistical advice to numerous business, nonbusiness, and academic communities. Her more recent involvement has been in the areas of education, medicine, and nonprofit capacity building.

Szabat received a B.S. in mathematics from State University of New York at Albany and M.S. and Ph.D. degrees in statistics, with a cognate in operations research, from the Wharton School of the University of Pennsylvania.

Brief Contents

Preface xix

Getting Started: Important Things to Learn First 1

- 1** Defining and Collecting Data 13
- 2** Organizing and Visualizing Variables 36
- 3** Numerical Descriptive Measures 101
- 4** Basic Probability 151
- 5** Discrete Probability Distributions 185
- 6** The Normal Distribution and Other Continuous Distributions 219
- 7** Sampling Distributions 250
- 8** Confidence Interval Estimation 272
- 9** Fundamentals of Hypothesis Testing: One-Sample Tests 308
- 10** Two-Sample Tests 347
- 11** Analysis of Variance 394
- 12** Chi-Square and Nonparametric Tests 447
- 13** Simple Linear Regression 491
- 14** Introduction to Multiple Regression 543
- 15** Multiple Regression Model Building 596
- 16** Time-Series Forecasting 629
- 17** Business Analytics 674
- 18** A Roadmap for Analyzing Data 707
- 19** Statistical Applications in Quality Management (*online*)
- 20** Decision Making (*online*)

Appendices A–G 715

Self-Test Solutions and Answers to Selected Even-Numbered Problems 767

Index 803

Contents

Preface xix

Getting Started: Important Things to Learn First 1

USING STATISTICS: “You Cannot Escape from Data” 1

GS.1 Statistics: A Way of Thinking 2

GS.2 Data: What Is It? 3

GS.3 Business Analytics: The Changing Face of Statistics 4
“Big Data” 4
Statistics: An Important Part of Your Business Education 5

GS.4 Software and Statistics 6
Excel and Minitab Guides 6

REFERENCES 7

KEY TERMS 7

EXCEL GUIDE 8

EG1. Getting Started with Microsoft Excel 8

EG2. Entering Data 8

EG3. Opening and Saving Workbooks 9

EG4. Creating and Copying Worksheets 10

EG5. Printing Worksheets 10

MINITAB GUIDE 11

MG1. Getting Started with Minitab 11

MG2. Entering Data 11

MG3. Opening and Saving Worksheets and Projects 11

MG4. Creating and Copying Worksheets 12

MG5. Printing Parts of a Project 12

1 Defining and Collecting Data 13

USING STATISTICS: Beginning of the End ... Or the End of the Beginning? 13

1.1 Defining Data 14

Establishing the Variable Type 14

1.2 Measurement Scales for Variables 15

Nominal and Ordinal Scales 15
Interval and Ratio Scales 16

1.3 Collecting Data 18

Data Sources 18
Populations and Samples 19
Data Formatting 19
Data Cleaning 20
Recoding Variables 20

1.4 Types of Sampling Methods 21

Simple Random Sample 22
Systematic Sample 23

Stratified Sample 23

Cluster Sample 23

1.5 Types of Survey Errors 24

Coverage Error 25

Nonresponse Error 25

Sampling Error 25

Measurement Error 25

Ethical Issues About Surveys 26

THINK ABOUT THIS: New Media Surveys/Old Sampling Problems 26

USING STATISTICS: Beginning of the End ... Revisited 27

SUMMARY 28

REFERENCES 28

KEY TERMS 28

CHECKING YOUR UNDERSTANDING 29

CHAPTER REVIEW PROBLEMS 29

CASES FOR CHAPTER 1 30

Managing Ashland MultiComm Services 30

CardioGood Fitness 30

Clear Mountain State Student Surveys 31

Learning with the Digital Cases 31

CHAPTER 1 EXCEL GUIDE 33

EG1.1 Defining Data 33

EG1.2 Measurement Scales for Variables 33

EG1.3 Collecting Data 33

EG1.4 Types of Sampling Methods 33

CHAPTER 1 MINITAB GUIDE 34

MG1.1 Defining Data 34

MG1.2 Measurement Scales for Variables 34

MG1.3 Collecting Data 35

MG1.4 Types of Sampling Methods 35

2 Organizing and Visualizing Variables 36

USING STATISTICS: The Choice Is Yours 36

How to Proceed with This Chapter 37

2.1 Organizing Categorical Variables 38

The Summary Table 38

The Contingency Table 39

2.2 Organizing Numerical Variables 42

The Ordered Array 42

The Frequency Distribution 43

Classes and Excel Bins 45

The Relative Frequency Distribution and the Percentage Distribution 45

The Cumulative Distribution 47

Stacked and Unstacked Data 49

2.3 Visualizing Categorical Variables 51	3.2 Variation and Shape 107
The Bar Chart 51	The Range 107
The Pie Chart 52	The Variance and the Standard Deviation 108
The Pareto Chart 53	The Coefficient of Variation 112
The Side-by-Side Bar Chart 55	Z Scores 113
2.4 Visualizing Numerical Variables 57	Shape: Skewness and Kurtosis 114
The Stem-and-Leaf Display 57	VISUAL EXPLORATIONS: Exploring Descriptive Statistics 117
The Histogram 59	3.3 Exploring Numerical Data 120
The Percentage Polygon 60	Quartiles 120
The Cumulative Percentage Polygon (Ogive) 61	The Interquartile Range 122
2.5 Visualizing Two Numerical Variables 65	The Five-Number Summary 123
The Scatter Plot 65	The Boxplot 124
The Time-Series Plot 66	3.4 Numerical Descriptive Measures for a Population 127
2.6 Organizing Many Categorical Variables 68	The Population Mean 127
2.7 Challenges in Organizing and Visualizing Variables 70	The Population Variance and Standard Deviation 128
Obscuring Data 70	The Empirical Rule 129
Creating False Impressions 71	The Chebyshev Rule 130
Chartjunk 72	3.5 The Covariance and the Coefficient of Correlation 131
Guidelines for Constructing Visualizations 74	The Covariance 132
USING STATISTICS: The Choice <i>Is Yours</i> , Revisited 75	The Coefficient of Correlation 133
SUMMARY 75	3.6 Descriptive Statistics: Pitfalls and Ethical Issues 137
REFERENCES 76	USING STATISTICS: More Descriptive Choices, Revisited 138
KEY EQUATIONS 76	SUMMARY 138
KEY TERMS 77	REFERENCES 139
CHECKING YOUR UNDERSTANDING 77	KEY EQUATIONS 139
CHAPTER REVIEW PROBLEMS 77	KEY TERMS 140
CASES FOR CHAPTER 2 82	CHECKING YOUR UNDERSTANDING 140
Managing Ashland MultiComm Services 82	CHAPTER REVIEW PROBLEMS 141
Digital Case 83	CASES FOR CHAPTER 3 144
CardioGood Fitness 83	Managing Ashland MultiComm Services 144
The Choice <i>Is Yours</i> Follow-Up 83	Digital Case 144
Clear Mountain State Student Surveys 84	CardioGood Fitness 144
CHAPTER 2 EXCEL GUIDE 85	More Descriptive Choices Follow-up 144
EG2.1 Organizing Categorical Variables 85	Clear Mountain State Student Surveys 144
EG2.2 Organizing Numerical Variables 87	CHAPTER 3 EXCEL GUIDE 145
EG2.3 Visualizing Categorical Variables 89	EG3.1 Central Tendency 145
EG2.4 Visualizing Numerical Variables 91	EG3.2 Variation and Shape 145
EG2.5 Visualizing Two Numerical Variables 94	EG3.3 Exploring Numerical Data 146
EG2.6 Organizing Many Categorical Variables 94	EG3.4 Numerical Descriptive Measures for a Population 147
CHAPTER 2 MINITAB GUIDE 95	EG3.5 The Covariance and the Coefficient of Correlation 147
MG2.1 Organizing Categorical Variables 95	CHAPTER 3 MINITAB GUIDE 148
MG2.2 Organizing Numerical Variables 96	MG3.1 Central Tendency 148
MG2.3 Visualizing Categorical Variables 96	MG3.2 Variation and Shape 148
MG2.4 Visualizing Numerical Variables 98	MG3.3 Exploring Numerical Data 149
MG2.5 Visualizing Two Numerical Variables 100	MG3.4 Numerical Descriptive Measures for a Population 149
MG2.6 Organizing Many Categorical Variables 100	MG3.5 The Covariance and the Coefficient of Correlation 149

3 Numerical Descriptive Measures

101

USING STATISTICS: More Descriptive Choices 101

- 3.1 Central Tendency 102
 - The Mean 102
 - The Median 104
 - The Mode 105
 - The Geometric Mean 106

4 Basic Probability

151

USING STATISTICS: Possibilities at M&R Electronics World 151

- 4.1 Basic Probability Concepts 152
 - Events and Sample Spaces 153
 - Contingency Tables and Venn Diagrams 155
 - Simple Probability 155
 - Joint Probability 156

Marginal Probability 157	5.4 Poisson Distribution 202
General Addition Rule 158	5.5 Hypergeometric Distribution 206
4.2 Conditional Probability 161	5.6 Using the Poisson Distribution to Approximate the Binomial Distribution (<i>online</i>) 209
Computing Conditional Probabilities 161	
Decision Trees 163	USING STATISTICS: Events of Interest at Ricknel Homecenters, Revisited 209
Independence 165	SUMMARY 209
Multiplication Rules 166	REFERENCES 209
Marginal Probability Using the General Multiplication Rule 167	KEY EQUATIONS 210
4.3 Bayes' Theorem 169	KEY TERMS 210
THINK ABOUT THIS: Divine Providence and Spam 172	CHECKING YOUR UNDERSTANDING 211
4.4 Counting Rules 174	CHAPTER REVIEW PROBLEMS 211
4.5 Ethical Issues and Probability 177	
USING STATISTICS: Possibilities at M&R Electronics World, Revisited 178	CASES FOR CHAPTER 5 213
SUMMARY 178	Managing Ashland MultiComm Services 213
REFERENCES 178	Digital Case 214
KEY EQUATIONS 179	
KEY TERMS 179	CHAPTER 5 EXCEL GUIDE 215
CHECKING YOUR UNDERSTANDING 180	EG5.1 The Probability Distribution for a Discrete Variable 215
CHAPTER REVIEW PROBLEMS 180	EG5.2 Covariance of a Probability Distribution and its Application in Finance 215
CASES FOR CHAPTER 4 182	EG5.3 Binomial Distribution 215
Digital Case 182	EG5.4 Poisson Distribution 216
CardioGood Fitness 182	EG5.5 Hypgeometric Distribution 216
The Choice Is Yours Follow-Up 182	
Clear Mountain State Student Surveys 182	CHAPTER 5 MINITAB GUIDE 217
CHAPTER 4 EXCEL GUIDE 183	MG5.1 The Probability Distribution for a Discrete Variable 217
EG4.1 Basic Probability Concepts 183	MG5.2 Covariance and its Application in Finance 217
EG4.2 Conditional Probability 183	MG5.3 Binomial Distribution 217
EG4.3 Bayes' Theorem 183	MG5.4 Poisson Distribution 217
EG4.4 Counting Rules 183	MG5.5 Hypergeometric Distribution 218
CHAPTER 4 MINITAB GUIDE 184	
MG4.1 Basic Probability Concepts 184	
MG4.2 Conditional Probability 184	
MG4.3 Bayes' Theorem 184	
MG4.4 Counting Rules 184	

5 Discrete Probability Distributions 185

USING STATISTICS: Events of Interest at Ricknel Home Centers 185

5.1 The Probability Distribution for a Discrete Variable 186	
Expected Value of a Discrete Variable 186	
Variance and Standard Deviation of a Discrete Variable 187	
5.2 Covariance of a Probability Distribution and Its Application in Finance 189	
Covariance 190	
Expected Value, Variance, and Standard Deviation of the Sum of Two Variables 191	
Portfolio Expected Return and Portfolio Risk 191	
5.3 Binomial Distribution 195	

6 The Normal Distribution and Other Continuous Distributions 219

USING STATISTICS: Normal Downloading at MyTVLab 219

6.1 Continuous Probability Distributions 220	
6.2 The Normal Distribution 220	
Computing Normal Probabilities 222	
Finding X Values 227	
VISUAL EXPLORATIONS: Exploring the Normal Distribution 231	
THINK ABOUT THIS: What Is Normal? 231	
6.3 Evaluating Normality 233	
Comparing Data Characteristics to Theoretical Properties 233	
Constructing the Normal Probability Plot 235	
6.4 The Uniform Distribution 237	
6.5 The Exponential Distribution 240	
6.6 The Normal Approximation to the Binomial Distribution (<i>online</i>) 242	

USING STATISTICS: Normal Downloading at MyTVLab, Revisited 242

SUMMARY	243	MG7.2 Sampling Distribution of the Mean	271	
REFERENCES	243	MG7.3 Sampling Distribution of the Proportion	271	
KEY EQUATIONS	243	<hr/>		
KEY TERMS	244	8 Confidence Interval Estimation		
CHECKING YOUR UNDERSTANDING	244	272		
CHAPTER REVIEW PROBLEMS	244	<hr/>		
CASES FOR CHAPTER 6	245	USING STATISTICS: Getting Estimates at Ricknel Home Centers		
Managing Ashland MultiComm Services	245	272		
Digital Case	246	8.1 Confidence Interval Estimate for the Mean (σ Known)		
CardioGood Fitness	246	273 Can You Ever Know the Population Standard Deviation?		
More Descriptive Choices Follow-up	246	8.2 Confidence Interval Estimate for the Mean (σ Unknown)		
Clear Mountain State Student Surveys	246	279 Student's <i>t</i> Distribution Properties of the <i>t</i> Distribution The Concept of Degrees of Freedom The Confidence Interval Statement		
CHAPTER 6 EXCEL GUIDE	247	8.3 Confidence Interval Estimate for the Proportion		
EG6.1 Continuous Probability Distributions	247	287		
EG6.2 The Normal Distribution	247	8.4 Determining Sample Size		
EG6.3 Evaluating Normality	247	290 Sample Size Determination for the Mean Sample Size Determination for the Proportion		
EG6.4 The Uniform Distribution	248	8.5 Confidence Interval Estimation and Ethical Issues		
EG6.5 The Exponential Distribution	248	295		
CHAPTER 6 MINITAB GUIDE	248	8.6 Application of Confidence Interval Estimation in Auditing (<i>online</i>)		
MG6.1 Continuous Probability Distributions	248	296		
MG6.2 The Normal Distribution	248	8.7 Estimation and Sample Size Estimation for Finite Populations (<i>online</i>)		
MG6.3 Evaluating Normality	248	296		
MG6.4 The Uniform Distribution	249	8.8 Bootstrapping (<i>online</i>)		
MG6.5 The Exponential Distribution	249	USING STATISTICS: Getting Estimates at Ricknel Home Centers, Revisited		
7 Sampling Distributions				296
USING STATISTICS: Sampling Oxford Cereals				250
7.1 Sampling Distributions	251	SUMMARY		
7.2 Sampling Distribution of the Mean	251	297		
The Unbiased Property of the Sample Mean	251	REFERENCES		
Standard Error of the Mean	253	KEY EQUATIONS		
Sampling from Normally Distributed Populations	254	KEY TERMS		
Sampling from Non-normally Distributed Populations—The Central Limit Theorem	257	CHECKING YOUR UNDERSTANDING		
VISUAL EXPLORATIONS: Exploring Sampling Distributions	261	CHAPTER REVIEW PROBLEMS		
7.3 Sampling Distribution of the Proportion	262	<hr/>		
7.4 Sampling from Finite Populations (<i>online</i>)	265	CASES FOR CHAPTER 8		
USING STATISTICS: Sampling Oxford Cereals, Revisited				301
SUMMARY	266	Managing Ashland MultiComm Services		
REFERENCES	266	Digital Case		
KEY EQUATIONS	266	Sure Value Convenience Stores		
KEY TERMS	267	CardioGood Fitness		
CHECKING YOUR UNDERSTANDING	267	More Descriptive Choices Follow-Up		
CHAPTER REVIEW PROBLEMS	267	Clear Mountain State Student Surveys		
CASES FOR CHAPTER 7				CHAPTER 8 EXCEL GUIDE
Managing Ashland MultiComm Services	269	304		
Digital Case	269	EG8.1 Confidence Interval Estimate for the Mean (σ Known)		
CHAPTER 7 EXCEL GUIDE				304
EG7.1 Sampling Distributions	270	EG8.2 Confidence Interval Estimate for the Mean (σ Unknown)		
EG7.2 Sampling Distribution of the Mean	270	305		
EG7.3 Sampling Distribution of the Proportion	270	EG8.3 Confidence Interval Estimate for the Proportion		
CHAPTER 7 MINITAB GUIDE				305
MG7.1 Sampling Distributions	271	EG8.4 Determining Sample Size		

9 Fundamentals of Hypothesis Testing: One-Sample Tests 308

USING STATISTICS: Significant Testing at Oxford Cereals 308

- 9.1 Fundamentals of Hypothesis-Testing Methodology 309
 - The Null and Alternative Hypotheses 309
 - The Critical Value of the Test Statistic 310
 - Regions of Rejection and Nonrejection 311
 - Risks in Decision Making Using Hypothesis Testing 311
 - Z Test for the Mean (σ Known) 314
 - Hypothesis Testing Using the Critical Value Approach 314
 - Hypothesis Testing Using the *p*-Value Approach 317
 - A Connection Between Confidence Interval Estimation and Hypothesis Testing 319
 - Can You Ever Know the Population Standard Deviation? 320
- 9.2 *t* Test of Hypothesis for the Mean (σ Unknown) 321
 - The Critical Value Approach 322
 - The *p*-Value Approach 324
 - Checking the Normality Assumption 324
- 9.3 One-Tail Tests 328
 - The Critical Value Approach 328
 - The *p*-Value Approach 329
- 9.4 Z Test of Hypothesis for the Proportion 332
 - The Critical Value Approach 333
 - The *p*-Value Approach 334
- 9.5 Potential Hypothesis-Testing Pitfalls and Ethical Issues 336
 - Statistical Significance Versus Practical Significance 336
 - Statistical *Insignificance* Versus Importance 337
 - Reporting of Findings 337
 - Ethical Issues 337
- 9.6 Power of a Test (*online*) 337

USING STATISTICS: Significant Testing at Oxford Cereals, Revisited 338

- SUMMARY 338
- REFERENCES 338
- KEY EQUATIONS 339
- KEY TERMS 339
- CHECKING YOUR UNDERSTANDING 339
- CHAPTER REVIEW PROBLEMS 339

CASES FOR CHAPTER 9 341

- Managing Ashland MultiComm Services 341
- Digital Case 342
- Sure Value Convenience Stores 342

CHAPTER 9 EXCEL GUIDE 343

- EG9.1 Fundamentals of Hypothesis-Testing Methodology 343
- EG9.2 *t* Test of Hypothesis for the Mean (σ Unknown) 343
- EG9.3 One-Tail Tests 344
- EG9.4 Z Test of Hypothesis for the Proportion 344

CHAPTER 9 MINITAB GUIDE 345

- MG9.1 Fundamentals of Hypothesis-Testing Methodology 345
- MG9.2 *t* Test of Hypothesis for the Mean (σ Unknown) 345

MG9.3 One-Tail Tests 345
MG9.4 Z Test of Hypothesis for the Proportion 346

10 Two-Sample Tests 347

USING STATISTICS: For North Fork, Are There Different Means to the Ends? 347

- 10.1 Comparing the Means of Two Independent Populations 348
 - Pooled-Variance *t* Test for the Difference Between Two Means 348
 - Confidence Interval Estimate for the Difference Between Two Means 353
 - t* Test for the Difference Between Two Means, Assuming Unequal Variances 354
 - Do People Really Do This? 356
 - 10.2 Comparing the Means of Two Related Populations 359
 - Paired *t* Test 360
 - Confidence Interval Estimate for the Mean Difference 365
 - 10.3 Comparing the Proportions of Two Independent Populations 367
 - Z Test for the Difference Between Two Proportions 367
 - Confidence Interval Estimate for the Difference Between Two Proportions 371
 - 10.4 *F* Test for the Ratio of Two Variances 373
 - 10.5 Effect Size (*online*)
- USING STATISTICS:** For North Fork, Are There Different Means to the Ends? Revisited 378
- SUMMARY 379
 - REFERENCES 380
 - KEY EQUATIONS 380
 - KEY TERMS 381
 - CHECKING YOUR UNDERSTANDING 381
 - CHAPTER REVIEW PROBLEMS 381

CASES FOR CHAPTER 10 383

- Managing Ashland MultiComm Services 383
- Digital Case 384
- Sure Value Convenience Stores 384
- CardioGood Fitness 384
- More Descriptive Choices Follow-Up 385
- Clear Mountain State Student Surveys 385

CHAPTER 10 EXCEL GUIDE 386

- EG10.1 Comparing the Means of Two Independent Populations 386
- EG10.2 Comparing the Means of Two Related Populations 388
- EG10.3 Comparing the Proportions of Two Independent Populations 389
- EG10.4 *F* Test for the Ratio of Two Variances 389

CHAPTER 10 MINITAB GUIDE 391

- MG10.1 Comparing the Means of Two Independent Populations 391
- MG10.2 Comparing the Means of Two Related Populations 391
- MG10.3 Comparing the Proportions of Two Independent Populations 392
- MG10.4 *F* Test for the Ratio of Two Variances 392

11 Analysis of Variance 394

USING STATISTICS: The Means to Find Differences at Arlington's 394

- 11.1 The Completely Randomized Design: One-Way ANOVA 395
 - Analyzing Variation in One-Way ANOVA 396
 - F* Test for Differences Among More Than Two Means 398
 - Multiple Comparisons: The Tukey-Kramer Procedure 402
 - The Analysis of Means (ANOM) (*online*) 404
 - ANOVA Assumptions 405
 - Levene Test for Homogeneity of Variance 405
- 11.2 The Randomized Block Design 410
 - Testing for Factor and Block Effects 410
 - Multiple Comparisons: The Tukey Procedure 415
- 11.3 The Factorial Design: Two-Way ANOVA 418
 - Factor and Interaction Effects 419
 - Testing for Factor and Interaction Effects 421
 - Multiple Comparisons: The Tukey Procedure 424
 - Visualizing Interaction Effects: The Cell Means Plot 426
 - Interpreting Interaction Effects 426
- 11.4 Fixed Effects, Random Effects, and Mixed Effects Models (*online*) 431

USING STATISTICS: The Means to Find Differences at Arlington's Revisited 431

- SUMMARY** 431
- REFERENCES** 432
- KEY EQUATIONS** 432
- KEY TERMS** 433
- CHECKING YOUR UNDERSTANDING** 434
- CHAPTER REVIEW PROBLEMS** 434

CASES FOR CHAPTER 11 437

- Managing Ashland MultiComm Services 437
- Digital Case 437
- Sure Value Convenience Stores 438
- CardioGood Fitness 438
- More Descriptive Choices Follow-Up 438
- Clear Mountain State Student Surveys 438

CHAPTER 11 EXCEL GUIDE 440

- EG11.1 The Completely Randomized Design: One-Way ANOVA 440
- EG11.2 The Randomized Block Design 442
- EG11.3 The Factorial Design: Two-Way ANOVA 443

CHAPTER 11 MINITAB GUIDE 444

- MG11.1 The Completely Randomized Design: One-Way ANOVA 444
- MG11.2 The Randomized Block Design 445
- MG11.3 The Factorial Design: Two-Way ANOVA 445

12 Chi-Square and Nonparametric Tests 447

USING STATISTICS: Avoiding Guesswork About Resort Guests 447

- 12.1 Chi-Square Test for the Difference Between Two Proportions 448

- 12.2 Chi-Square Test for Differences Among More Than Two Proportions 455
 - The Marascuilo Procedure 458
 - The Analysis of Proportions (ANOP) (*online*) 460
- 12.3 Chi-Square Test of Independence 461
- 12.4 Wilcoxon Rank Sum Test: A Nonparametric Method for Two Independent Populations 467
- 12.5 Kruskal-Wallis Rank Test: A Nonparametric Method for the One-Way ANOVA 473
 - Assumptions 476
- 12.6 McNemar Test for the Difference Between Two Proportions (Related Samples) (*online*) 477
- 12.7 Chi-Square Test for the Variance or Standard Deviation (*online*) 478
- 12.8 Wilcoxon Signed Ranks Test: A Nonparametric Test for Two Related Populations (*online*) 478
- 12.9 Friedman Rank Test: A Nonparametric Test for the Randomized Block Design (*online*) 478

USING STATISTICS: Avoiding Guesswork About Resort Guests, Revisited 478

- SUMMARY** 479
- REFERENCES** 479
- KEY EQUATIONS** 480
- KEY TERMS** 480
- CHECKING YOUR UNDERSTANDING** 480
- CHAPTER REVIEW PROBLEMS** 480

CASES FOR CHAPTER 12 482

- Managing Ashland MultiComm Services 482
- Digital Case 483
- Sure Value Convenience Stores 483
- CardioGood Fitness 484
- More Descriptive Choices Follow-Up 484
- Clear Mountain State Student Surveys 484

CHAPTER 12 EXCEL GUIDE 486

- EG12.1 Chi-Square Test for the Difference Between Two Proportions 486
- EG12.2 Chi-Square Test for Differences Among More Than Two Proportions 486
- EG12.3 Chi-Square Test of Independence 487
- EG12.4 Wilcoxon Rank Sum Test: a Nonparametric Method for Two Independent Populations 487
- EG12.5 Kruskal-Wallis Rank Test: a Nonparametric Method for the One-Way ANOVA 488

CHAPTER 12 MINITAB GUIDE 489

- MG12.1 Chi-Square Test for the Difference Between Two Proportions 489
- MG12.2 Chi-Square Test for Differences Among More Than Two Proportions 489
- MG12.3 Chi-Square Test of Independence 489
- MG12.4 Wilcoxon Rank Sum Test: a Nonparametric Method for Two Independent Populations 489
- MG12.5 Kruskal-Wallis Rank Test: a Nonparametric Method for the One-Way ANOVA 490

13 Simple Linear Regression 491

USING STATISTICS: Knowing Customers at Sunflowers Apparel 491

- 13.1 Types of Regression Models 492
 - Simple Linear Regression Models 493
- 13.2 Determining the Simple Linear Regression Equation 494
 - The Least-Squares Method 494
 - Predictions in Regression Analysis: Interpolation Versus Extrapolation 497
 - Computing the Y Intercept, b_0 , and the Slope, b_1 497
- 13.3 Measures of Variation 502
 - Computing the Sum of Squares 502
 - The Coefficient of Determination 503
 - Standard Error of the Estimate 505
- 13.4 Assumptions of Regression 507
- 13.5 Residual Analysis 507
 - Evaluating the Assumptions 507
- 13.6 Measuring Autocorrelation: The Durbin-Watson Statistic 511
 - Residual Plots to Detect Autocorrelation 511
 - The Durbin-Watson Statistic 512
- 13.7 Inferences About the Slope and Correlation Coefficient 515
 - t Test for the Slope 516
 - F Test for the Slope 517
 - Confidence Interval Estimate for the Slope 519
 - t Test for the Correlation Coefficient 519
- 13.8 Estimation of Mean Values and Prediction of Individual Values 523
 - The Confidence Interval Estimate for the Mean Response 523
 - The Prediction Interval for an Individual Response 524
- 13.9 Potential Pitfalls in Regression 527
 - Six Steps for Avoiding the Potential Pitfalls 529

USING STATISTICS: Knowing Customers at Sunflowers Apparel, Revisited 529

- SUMMARY 529
- REFERENCES 530
- KEY EQUATIONS 531
- KEY TERMS 532
- CHECKING YOUR UNDERSTANDING 532
- CHAPTER REVIEW PROBLEMS 532

CASES FOR CHAPTER 13 536

- Managing Ashland MultiComm Services 536
- Digital Case 536
- Brynne Packaging 536

CHAPTER 13 EXCEL GUIDE 538

- EG13.1 Types of Regression Models 538
- EG13.2 Determining the Simple Linear Regression Equation 538
- EG13.3 Measures of Variation 539
- EG13.4 Assumptions of Regression 539
- EG13.5 Residual Analysis 539
- EG13.6 Measuring Autocorrelation: the Durbin-Watson Statistic 540

- EG13.7 Inferences About the Slope and Correlation Coefficient 540
- EG13.8 Estimation of Mean Values and Prediction of Individual Values 540

CHAPTER 13 MINITAB GUIDE 541

- MG13.1 Types of Regression Models 541
- MG13.2 Determining the Simple Linear Regression Equation 541
- MG13.3 Measures of Variation 541
- MG13.4 Assumptions 541
- MG13.5 Residual Analysis 541
- MG13.6 Measuring Autocorrelation: the Durbin-Watson Statistic 542
- MG13.7 Inferences About the Slope and Correlation Coefficient 542
- MG13.8 Estimation of Mean Values and Prediction of Individual Values 542

14 Introduction to Multiple Regression 543

USING STATISTICS: The Multiple Effects of OmniPower Bars 543

- 14.1 Developing a Multiple Regression Model 544
 - Interpreting the Regression Coefficients 545
 - Predicting the Dependent Variable Y 547
- 14.2 r^2 , Adjusted r^2 , and the Overall F Test 550
 - Coefficient of Multiple Determination 550
 - Adjusted r^2 550
 - Test for the Significance of the Overall Multiple Regression Model 551
- 14.3 Residual Analysis for the Multiple Regression Model 553
- 14.4 Inferences Concerning the Population Regression Coefficients 555
 - Tests of Hypothesis 555
 - Confidence Interval Estimation 556
- 14.5 Testing Portions of the Multiple Regression Model 558
 - Coefficients of Partial Determination 562
- 14.6 Using Dummy Variables and Interaction Terms in Regression Models 563
 - Dummy Variables 564
 - Interactions 566
- 14.7 Logistic Regression 573
- 14.8 Influence Analysis 578
 - The Hat Matrix Elements, h_i 579
 - The Studentized Deleted Residuals, t_i 579
 - Cook's Distance Statistic, D_i 579
 - Comparison of Statistics 580

USING STATISTICS: The Multiple Effects of Omnipower Bars, Revisited 581

- SUMMARY 581
- REFERENCES 583
- KEY EQUATIONS 583
- KEY TERMS 584
- CHECKING YOUR UNDERSTANDING 584
- CHAPTER REVIEW PROBLEMS 584

CASES FOR CHAPTER 14 587

- Managing Ashland MultiComm Services 587
- Digital Case 588

CHAPTER 14 EXCEL GUIDE 589

- EG14.1 Developing a Multiple Regression Model 589
- EG14.2 r^2 , Adjusted r^2 , and the Overall F Test 590
- EG14.3 Residual Analysis for the Multiple Regression Model 590
- EG14.4 Inferences Concerning the Population Regression Coefficients 591
- EG14.5 Testing Portions of the Multiple Regression Model 591
- EG14.6 Using Dummy Variables and Interaction Terms in Regression Models 591
- EG14.7 Logistic Regression 591
- EG14.8 Influence Analysis 592

CHAPTER 14 MINITAB GUIDE 592

- MG14.1 Developing a Multiple Regression Model 592
- MG14.2 r^2 , Adjusted r^2 , and the Overall F Test 593
- MG14.3 Residual Analysis for the Multiple Regression Model 593
- MG14.4 Inferences Concerning the Population Regression Coefficients 593
- MG14.5 Testing Portions of the Multiple Regression Model 594
- MG14.6 Using Dummy Variables and Interaction Terms in Regression Models 594
- MG14.7 Logistic Regression 594
- MG14.8 Influence Analysis 595

15 Multiple Regression Model Building 596

USING STATISTICS: Valuing Parsimony at WSTA-TV 596

- 15.1 The Quadratic Regression Model 597
 - Finding the Regression Coefficients and Predicting Y 597
 - Testing for the Significance of the Quadratic Model 599
 - Testing the Quadratic Effect 600
 - The Coefficient of Multiple Determination 602
- 15.2 Using Transformations in Regression Models 604
 - The Square-Root Transformation 605
 - The Log Transformation 605
- 15.3 Collinearity 608
- 15.4 Model Building 609
 - The Stepwise Regression Approach to Model Building 611
 - The Best-Subsets Approach to Model Building 612
 - Model Validation 616
 - Steps for Successful Model Building 616
- 15.5 Pitfalls in Multiple Regression and Ethical Issues 618
 - Pitfalls in Multiple Regression 618
 - Ethical Issues 618

USING STATISTICS: Valuing Parsimony at WSTA-TV, Revisited 619

- SUMMARY 619
- KEY EQUATIONS 619
- REFERENCES 621
- KEY TERMS 621
- CHECKING YOUR UNDERSTANDING 621
- CHAPTER REVIEW PROBLEMS 621

CASES FOR CHAPTER 15 623

- The Mountain States Potato Company 623
- Sure Value Convenience Stores 623
- Digital Case 623

The Craybill Instrumentation Company Case 624
More Descriptive Choices Follow-Up 624

CHAPTER 15 EXCEL GUIDE 625

- EG15.1 The Quadratic Regression Model 625
- EG15.2 Using Transformations in Regression Models 625
- EG15.3 Collinearity 625
- EG15.4 Model Building 626

CHAPTER 15 MINITAB GUIDE 626

- MG15.1 The Quadratic Regression Model 626
- MG15.2 Using Transformations in Regression Models 627
- MG15.3 Collinearity 627
- MG15.4 Model Building 627

16 Time-Series Forecasting 629

USING STATISTICS: Principled Forecasting 629

- 16.1 The Importance of Business Forecasting 630
- 16.2 Component Factors of Time-Series Models 630
- 16.3 Smoothing an Annual Time Series 631
 - Moving Averages 632
 - Exponential Smoothing 634
- 16.4 Least-Squares Trend Fitting and Forecasting 637
 - The Linear Trend Model 637
 - The Quadratic Trend Model 639
 - The Exponential Trend Model 641
 - Model Selection Using First, Second, and Percentage Differences 643
- 16.5 Autoregressive Modeling for Trend Fitting and Forecasting 647
 - Selecting an Appropriate Autoregressive Model 648
 - Determining the Appropriateness of a Selected Model 649
- 16.6 Choosing an Appropriate Forecasting Model 655
 - Performing a Residual Analysis 655
 - Measuring the Magnitude of the Residuals Through Squared or Absolute Differences 655
 - Using the Principle of Parsimony 656
 - A Comparison of Four Forecasting Methods 656
- 16.7 Time-Series Forecasting of Seasonal Data 658
 - Least-Squares Forecasting with Monthly or Quarterly Data 659
- 16.8 Index Numbers (*online*) 664

THINK ABOUT THIS: Let the Model User Beware 664**USING STATISTICS: Principled Forecasting, Revisited 664**

- SUMMARY 665
- REFERENCES 665
- KEY EQUATIONS 666
- KEY TERMS 666
- CHECKING YOUR UNDERSTANDING 667
- CHAPTER REVIEW PROBLEMS 667

CASES FOR CHAPTER 16 668

- Managing Ashland MultiComm Services 668
- Digital Case 668

CHAPTER 16 EXCEL GUIDE 669

- EG16.1 The Importance of Business Forecasting 669
- EG16.2 Component Factors of Time-Series Models 669
- EG16.3 Smoothing an Annual Time Series 669

- EG16.4 Least-Squares Trend Fitting and Forecasting 670
 EG16.5 Autoregressive Modeling for Trend Fitting and Forecasting 670
 EG16.6 Choosing an Appropriate Forecasting Model 671
 EG16.7 Time-Series Forecasting of Seasonal Data 671

CHAPTER 16 MINITAB GUIDE 672

- MG16.1 The Importance of Business Forecasting 672
 MG16.2 Component Factors of Time-Series Models 672
 MG16.3 Smoothing an Annual Time Series 672
 MG16.4 Least-Squares Trend Fitting and Forecasting 673
 MG16.5 Autoregressive Modeling for Trend Fitting and Forecasting 673
 MG16.6 Choosing an Appropriate Forecasting Model 673
 MG16.7 Time-Series Forecasting of Seasonal Data 673

17 Business Analytics 674**USING STATISTICS:** Finding the Right Lines at WaldoLands 674

- 17.1 Descriptive Analytics 675
 Dashboards 676
 Data Discovery 678
 17.2 Predictive Analytics 682
 17.3 Classification and Regression Trees 683
 Regression Tree Example 685
 17.4 Neural Networks 688
 Multilayer Perceptrons 688
 17.5 Cluster Analysis 691
 17.6 Multidimensional Scaling 693

USING STATISTICS: Finding the Right Lines at Waldolands, Revisited 696

- REFERENCES 697
 KEY EQUATIONS 697
 KEY TERMS 697
 CHECKING YOUR UNDERSTANDING 698
 CHAPTER REVIEW PROBLEMS 698

CASE FOR CHAPTER 17

- The Mountain States Potato Company 699

CHAPTER 17 SOFTWARE GUIDE 700

- SG17.1 Descriptive Analytics 700
 SG17.2 Predictive Analytics 704
 SG17.3 Classification and Regression Trees 704
 SG17.4 Neural Networks 705
 SG17.5 Cluster Analysis 706
 SG17.6 Multidimensional Scaling 706

18 A Roadmap for Analyzing Data 707**USING STATISTICS:** Mounting Future Analyses 707

- 18.1 Analyzing Numerical Variables 709
 Describing the Characteristics of a Numerical Variable 710
 Reaching Conclusions About the Population Mean and/or Standard Deviation 710

- Determining Whether the Mean and/or Standard Deviation Differs Depending on the Group 710
 Determining Which Factors Affect the Value of a Variable 711
 Predicting the Value of a Variable Based on the Values of Other Variables 711
 Determining Whether the Values of a Variable Are Stable Over Time 711

- 18.2 Analyzing Categorical Variables 711
 Describing the Proportion of Items of Interest in Each Category 712
 Reaching Conclusions About the Proportion of Items of Interest 712
 Determining Whether the Proportion of Items of Interest Differs Depending on the Group 712
 Predicting the Proportion of Items of Interest Based on the Values of Other Variables 712
 Determining Whether the Proportion of Items of Interest Is Stable Over Time 712

USING STATISTICS: Mounting Future Analyses, Revisited 713
 Digital Case 713**CHAPTER REVIEW PROBLEMS 713****19 Statistical Applications in Quality Management (online)****USING STATISTICS:** Finding Quality at the Beachcomber

- 19.1 The Theory of Control Charts
 19.2 Control Chart for the Proportion: The *p* Chart
 19.3 The Red Bead Experiment: Understanding Process Variability
 19.4 Control Chart for an Area of Opportunity: The *c* Chart
 19.5 Control Charts for the Range and the Mean
 The *R* Chart
 The \bar{X} Chart
 19.6 Process Capability
 Customer Satisfaction and Specification Limits
 Capability Indices
 CPL , CPU , and C_{pk}
 19.7 Total Quality Management
 19.8 Six Sigma
 The DMAIC Model
 Roles in a Six Sigma Organization
 Lean Six Sigma

USING STATISTICS: Finding Quality at the Beachcomber, Revisited

- SUMMARY
 REFERENCES
 KEY EQUATIONS
 KEY TERMS
 CHECKING YOUR UNDERSTANDING
 CHAPTER REVIEW PROBLEMS

CASES FOR CHAPTER 19

- The Harnswell Sewing Machine Company Case
- Managing Ashland Multicomm Services

CHAPTER 19 EXCEL GUIDE

- EG19.1 The Theory of Control Charts
- EG19.2 Control Chart for the Proportion: The p Chart
- EG19.3 The Red Bead Experiment: Understanding Process Variability
- EG19.4 Control Chart for an Area of Opportunity: The c Chart
- EG19.5 Control Charts for the Range and the Mean
- EG19.6 Process Capability

20 Decision Making (online)

USING STATISTICS: Reliable Decision Making

- 20.1 Payoff Tables and Decision Trees
- 20.2 Criteria for Decision Making
 - Maximax Payoff
 - Maximin Payoff
 - Expected Monetary Value
 - Expected Opportunity Loss
 - Return-to-Risk Ratio
- 20.3 Decision Making with Sample Information
- 20.4 Utility

THINK ABOUT THIS: Risky Business**USING STATISTICS: Reliable Decision-Making, Revisited**

- SUMMARY
- REFERENCES
- KEY EQUATIONS
- KEY TERMS
- CHAPTER REVIEW PROBLEMS

CHAPTER 20 EXCEL GUIDE

- EG20.1 Payoff Tables and Decision Trees
- EG20.2 Criteria for Decision Making

Appendices 715

- A. Basic Math Concepts and Symbols 716
 - A.1 Rules for Arithmetic Operations 716
 - A.2 Rules for Algebra: Exponents and Square Roots 716
 - A.3 Rules for Logarithms 717
 - A.4 Summation Notation 718
 - A.5 Statistical Symbols 721
 - A.6 Greek Alphabet 721
- B. Required Excel Skills 722
 - B.1 Worksheet Entries and References 722
 - B.2 Absolute and Relative Cell References 723
 - B.3 Entering Formulas into Worksheets 723
 - B.4 Pasting with Paste Special 724
 - B.5 Basic Worksheet Cell Formatting 724
 - B.6 Chart Formatting 726

- B.7 Selecting Cell Ranges for Charts 727
- B.8 Deleting the “Extra” Bar from a Histogram 727
- B.9 Creating Histograms for Discrete Probability Distributions 728
- C. Online Resources 729
 - C.1 About the Online Resources for This Book 729
 - C.2 Accessing the MyStatLab Course Online 729
 - C.3 Details of Downloadable Files 729
 - C.4 PHStat 737
- D. Configuring Microsoft Excel 738
 - D.1 Getting Microsoft Excel Ready for Use (ALL) 738
 - D.2 Getting PHStat Ready for Use (ALL) 739
 - D.3 Configuring Excel Security for Add-In Usage (WIN) 739
 - D.4 Opening PHStat (ALL) 740
 - D.5 Using a Visual Explorations Add-in Workbook (ALL) 741
 - D.6 Checking for the Presence of the Analysis ToolPak or Solver Add-Ins (ALL) 741
- E. Tables 742
 - E.1 Table of Random Numbers 742
 - E.2 The Cumulative Standardized Normal Distribution 744
 - E.3 Critical Values of t 746
 - E.4 Critical Values of χ^2 748
 - E.5 Critical Values of F 749
 - E.6 Lower and Upper Critical Values, T_1 , of the Wilcoxon Rank Sum Test 753
 - E.7 Critical Values of the Studentized Range, Q 754
 - E.8 Critical Values, d_L and d_U , of the Durbin–Watson Statistic, D (Critical Values Are One-Sided) 756
 - E.9 Control Chart Factors 757
 - E.10 The Standardized Normal Distribution 758
- F. Useful Excel Knowledge 759
 - F.1 Useful Keyboard Shortcuts 759
 - F.2 Verifying Formulas and Worksheets 760
 - F.3 New Function Names 760
 - F.4 Understanding the Nonstatistical Functions 762
- G. Software FAQs 764
 - G.1 PHStat FAQs 764
 - G.2 Microsoft Excel FAQs 765
 - G.3 FAQs for New Users of Microsoft Excel 2013 766
 - G.4 Minitab FAQs 766

Self-Test Solutions and Answers to Selected Even-Numbered Problems 767**Index 803**

Preface

Over a generation ago, advances in “data processing” led to new business opportunities as first centralized and then desktop computing proliferated. The Information Age was born. Computer science became much more than just an adjunct to a mathematics curriculum, and whole new fields of studies, such as computer information systems, emerged.

More recently, further advances in information technologies have combined with data analysis techniques to create new opportunities in what is more data *science* than data *processing* or *computer* science. The world of business statistics has grown larger, bumping into other disciplines. And, in a reprise of something that occurred a generation ago, new fields of study, this time with names such as informatics, data analytics, and decision science, have emerged.

This time of change makes what is taught in business statistics and how it is taught all the more critical. These new fields of study all share statistics as a foundation for further learning. We are accustomed to thinking about change, as seeking ways to continuously improve the teaching of business statistics have always guided our efforts. We actively participate in Decision Sciences Institute (DSI), American Statistical Association (ASA), and Making Statistics More Effective in Schools and Business (MSMESB) conferences. We use the ASA’s Guidelines for Assessment and Instruction (GAISE) reports and combine them with our experiences teaching business statistics to a diverse student body at several large universities.

What to teach and how to teach it are particularly significant questions to ask during a time of change. As an author team, we bring a unique collection of experiences that we believe helps us find the proper perspective in balancing the old and the new. Our two lead authors, Mark L. Berenson and David M. Levine, were the first educators to create a business statistics textbook that discussed using statistical software and incorporated “computer output” as illustrations—just the first of many teaching and curricular innovations in their many years of teaching business statistics. Our newest co-author, Kathryn A. Szabat, has provided statistical advice to various business and nonbusiness communities. Her background in statistics and operations research and her experiences interacting with professionals in practice have guided her, as departmental chair, in developing a new, interdisciplinary academic department, Business Systems and Analytics, in response to the technology- and data-driven changes in business today.

All three of us benefit from our many years teaching undergraduate business subjects and the diversity of interests and efforts of our past co-author, Timothy Krehbiel. We are pleased to offer the innovations and new content that are itemized starting on the next page. As in prior editions, we are guided by these key learning principles:

- Help students see the relevance of statistics to their own careers by providing examples drawn from the functional areas in which they may be specializing.
- Emphasize interpretation of statistical results over mathematical computation.
- Give students ample practice in understanding how to apply statistics to business.
- Familiarize students with how to use statistical software to assist business decision making.
- Provide clear instructions to students for using statistical applications.

Read more about these principles on page xxviii.

What's New and Innovative in This Edition?

This thirteenth edition of *Basic Business Statistics* contains both new and innovative features and content, while refining and extending the use of the DCOVA (**D**efine, **C**ollect, **O**rganize, **V**isualize, and **A**nalyze) framework, first introduced in the twelfth edition as an integrated approach for applying statistics to help solve business problems.

Innovations

Getting Started: Important Things to Learn First—In a time of change, you can never know exactly what knowledge and background students bring into an introductory business statistics classroom. Add that to the need to curb the fear factor about learning statistics that so many students begin with, and there's a lot to cover even before you teach your first statistical concept.

We created “Getting Started: Important Things to Learn First” to meet this challenge. This unit sets the context for explaining what statistics is (not what students may think!) while ensuring that all students share an understanding of the forces that make learning business statistics critically important today. Especially designed for instructors teaching with course management tools, including those teaching hybrid or online courses, “Getting Started” has been developed to be posted online or otherwise distributed before the first class section begins and is available for download as explained in Appendix C.

Student Tips—In-margin notes reinforce hard-to-master concepts and provide quick study tips for mastering important details.

Discussion of Business Analytics—“Getting Started: Important Things to Learn First” quickly defines *business analytics* and *big data* and notes how these things are changing the face of statistics.

This material serves as an introduction to the new “Business Analytics” chapter (Chapter 17). This new chapter begins with a scenario that uses the management of a theme park to introduce applications of business analytics. The chapter begins by discussing descriptive visualization methods used for general oversight and applies them to issues raised in the scenario. Using other examples, the chapter then discusses the predictive analytics methods classification and regression trees, neural nets, cluster analysis, and multidimensional scaling that are in common use today.

Because standard Microsoft Excel and Minitab offer little or no support for the methods discussed, the chapter uses results created using JMP, the interactive data analysis software from the SAS Institute, and Tableau Public, the Web-based data visualization tool from Tableau Software, where appropriate. For those interested, a special *Software Guide* located at the end of the chapter explains how to use these two programs (and Microsoft Excel) to construct the results shown in the chapter.

PHStat version 4—For Microsoft Excel users, this new version of the Pearson Education statistics add-in contains several new and enhanced procedures, simpler set up, and is compatible with both Microsoft Windows and (Mac) OS X Excel versions.

Chapter Short Takes Online PDF documents (available for download as explained in Appendix C) that supply additional insights or explanations to important statistical concepts or details about the results presented in this book.

Revised and Enhanced Content

New Continuing End-of-Chapter Cases—This thirteenth edition features several new end-of-chapter cases. New and recurring throughout the book is a case that concerns analysis of sales and marketing data for home fitness equipment (CardioGood Fitness), a case that concerns pricing decisions made by a retailer (Sure Value Convenience Stores), and the More Descriptive Choices Follow-Up case, which extends the use of the retirement funds sample first introduced in Chapter 2. Also recurring is the Clear Mountain State Student Surveys case, which uses data collected from surveys of undergraduate and graduate students to practice and reinforce statistical methods learned in various chapters. This case replaces end-of-chapter questions related to the student survey database in the previous edition. Joining the Mountain States Potato Company regression case of the previous edition are new cases in simple linear regression (Brynne Packaging) and multiple regression (The Craybill Instrumentation Company).

Many New Applied Examples and Problems—Many of the applied examples throughout this book use new problems or revised data. Approximately 44% of the problems are new to this edition. The ends-of-section and ends-of-chapter problem sets contain many new problems that use data from *The Wall Street Journal*, *USA Today*, and other sources.

Revised Using Statistics Scenarios—There are new or revised Using Statistics scenarios in five chapters.

Checklist for Preparing to Use Microsoft Excel or Minitab with This Book—Found in Section GS.4 of “Getting Started: Important Things to Learn First,” this checklist explains for students which skills they will need and where they will find information about those skills in the book.

Revised Appendices Keyed to the Preparing to Use Microsoft Excel Checklist—The revised Appendix B discusses the Excel skills that readers need to make best use of the *In-Depth Excel* instructions in this book. Appendix F presents useful Excel knowledge, including a discussion of the new worksheet function names that were introduced in Excel 2010. Appendix G presents FAQs about using Excel and Minitab with this book.

Configuring Microsoft Excel Appendix—This revised Appendix D discusses the procedures and practices that will help readers that use Microsoft Excel to avoid common technical problems that might otherwise arise as they learn business statistics with this book.

Distinctive Features

We have continued many of the traditions of past editions and have highlighted some of these features below.

Using Statistics Business Scenarios—Each chapter begins with a Using Statistics example that shows how statistics is used in the functional areas of business—accounting, finance, information systems, management, and marketing. Each scenario is used throughout the chapter to provide an applied context for the concepts. The chapter concludes with a Using Statistics, Revisited section that reinforces the statistical methods and applications discussed in each chapter.

Emphasis on Data Analysis and Interpretation of Excel and Minitab Results—We believe that the use of computer software is an integral part of learning statistics. Our focus emphasizes analyzing data by interpreting results while reducing emphasis on doing computations. For example, in the coverage of tables and charts in Chapter 2, the focus is on the interpretation of various charts and on when to use each chart. In our coverage of hypothesis testing in Chapters 9 through 12, and regression and multiple regression in Chapters 13 through 15, extensive computer results have been included so that the *p*-value approach can be emphasized.

Pedagogical Aids—An active writing style is used, with boxed numbered equations, set-off examples to provide reinforcement for learning concepts, student tips, problems divided into “Learning the Basics” and “Applying the Concepts,” key equations, and key terms.

Digital Cases—In the Digital Cases, available for download as explained in Appendix C, learners must examine interactive PDF documents to sift through various claims and information in order to discover the data most relevant to a business case scenario. Learners then determine whether the conclusions and claims are supported by the data. In doing so, learners discover and learn how to identify common misuses of statistical information. (Instructional tips for using the Digital Cases and solutions to the Digital Cases are included in the Instructor’s Solutions Manual.)

Answers—Most answers to the even-numbered exercises are included at the end of the book.

Flexibility Using Excel—For almost every statistical method discussed, this book presents more than one way of using Excel. Students can use *In-Depth Excel* instructions to directly work with worksheet solution details *or* they can use either the PHStat instructions *or* the Analysis ToolPak instructions to automate the creation of those worksheet solutions.

PHStat—PHStat is the Pearson Education statistics add-in that you use with Microsoft Excel to help build solutions to statistical problems. With PHStat, you fill in simple-to-use dialog boxes and watch as PHStat creates a worksheet solution for you. PHStat allows you to use the Microsoft Excel statistical functions without having to first learn advanced Excel techniques or worrying about building worksheets from scratch. As a student studying statistics, you can focus mainly on learning statistics and not worry about having to fully master Excel as well.

Unlike other programs, PHStat solutions are real worksheets that contain real Excel calculations (called formulas in Excel). You can examine the contents of worksheet solutions to learn the appropriate functions and calculations necessary to apply a particular statistical method. With most of these worksheet solutions, you can change worksheet data and immediately see how those changes affect the results. This book uses PHStat version 4 which includes over 60 procedures that create Excel worksheets and charts for these statistical methods:

Descriptive Statistics: boxplot, descriptive summary, dot scale diagram, frequency distribution, histogram & polygons, Pareto diagram, scatter plot, stem-and-leaf display, one-way tables & charts, and two-way tables & charts

Probability and probability distributions: simple & joint probabilities, normal probability plot, and binomial, exponential, hypergeometric, and Poisson probability distributions

Sampling: sampling distributions simulation

Confidence interval estimation: for the mean, sigma unknown; for the mean, sigma known, for the population variance, for the proportion, and for the total difference

Sample size determination: for the mean and the proportion

One-sample tests: Z test for the mean, sigma known; t test for the mean, sigma unknown; chi-square test for the variance; and Z test for the proportion

Two-sample tests (unsummarized data): pooled-variance t test, separate-variance t test, paired t test, F test for differences in two variances, and Wilcoxon rank sum test

Two-sample tests (summarized data): pooled-variance t test, separate-variance t test, paired t test, Z test for the differences in two means, F test for differences in two variances, chi-square test for differences in two proportions, Z test for the difference in two proportions, and McNemar test

Multiple-sample tests: chi-square test, Marascuilo procedure Kruskal-Wallis rank test, Levene test, one-way ANOVA, Tukey-Kramer procedure randomized block design, and two-way ANOVA with replication

Regression: simple linear regression, multiple regression, best subsets, stepwise regression, and logistic regression

Control charts: p chart, c chart, and R and $X\bar{r}$ charts.

Decision-making: covariance and portfolio management, expected monetary value, expected opportunity loss, and opportunity loss

Data preparation: stack and unstack data

See Appendix Section C.4 for more information about PHStat.

Visual Explorations—The series of Excel workbooks that allow students to interactively explore important statistical concepts in descriptive statistics, the normal distribution, sampling distributions, and regression analysis. For example, in descriptive statistics, students observe the effect of changes in the data on the mean, median, quartiles, and standard deviation. With the normal distribution, students see the effect of changes in the mean and standard deviation on the areas under the normal curve. In sampling distributions, students use simulation to explore the effect of sample size on a sampling distribution. In regression analysis, students have the opportunity to fit a line and observe how changes in the slope and intercept affect the goodness of fit. The Visual Explorations workbooks are available for download as explained in Appendix C. (See Appendix Section C.4 to learn more about the workbooks that comprise Visual Explorations.)

Chapter-by-Chapter Changes Made for This Edition

Besides the new and innovative content described in “What’s New and Innovative in This Edition?” the thirteenth edition of *Basic Business Statistics* contains the following specific changes to each chapter. Highlights of the changes to the individual chapters are as follows:

Getting Started: Important Things to Learn First—This all-new chapter includes new material on business analytics and introduces the DCOVA framework and a basic vocabulary of statistics, both of which were introduced in Chapter 1 of the twelfth edition.

Chapter 1—Collecting data has been relocated to this chapter from Section 2.1. Sampling methods and types of survey errors have been relocated from Sections 7.1 and 7.2. There is a new subsection on data cleaning. The CardioGood Fitness and Clear Mountain State Surveys cases are included.

Chapter 2—Section 2.1, “Data Collection,” has been moved to Chapter 1. The chapter uses a new data set that contains a sample of 316 mutual funds and a new set of restaurant cost data. The CardioGood Fitness, The Choice Is Yours Follow-up, and Clear Mountain State Surveys cases are included.

Chapter 3—For many examples, this chapter uses the new mutual funds data set that is introduced in Chapter 2. There is increased coverage of skewness and kurtosis. There is a new example on

computing descriptive measures from a population using “Dogs of the Dow.” The CardioGood Fitness, More Descriptive Choices Follow-up, and Clear Mountain State Surveys cases are included.

Chapter 4—The chapter example has been updated. There are new problems throughout the chapter. The CardioGood Fitness, The Choice *Is Yours* Follow-up, and Clear Mountain State Surveys cases are included.

Chapter 5—There is an additional example on applying probability distributions in finance, and there are many new problems throughout the chapter. The notation used has been made more consistent.

Chapter 6—This chapter has an updated Using Statistics scenario and some new problems. The CardioGood Fitness, More Descriptive Choices Follow-up, and Clear Mountain State Surveys cases are included.

Chapter 7—Sections 7.1 and 7.2 have been moved to Chapter 1. An additional example of sampling distributions from a larger population has been included.

Chapter 8—This chapter includes an updated Using Statistics scenario and new examples and exercises throughout the chapter. The Sure Value Convenience Stores, CardioGood Fitness, More Descriptive Choices Follow-up, and Clear Mountain State Surveys cases are included. The section “Applications of Confidence Interval Estimation in Auditing” has been moved online. There is an online section on bootstrapping.

Chapter 9—This chapter includes additional coverage of the pitfalls of hypothesis testing. The Sure Value Convenience Stores case is included.

Chapter 10—This chapter has an updated Using Statistics scenario, a new example on the paired *t*-test on textbook prices, and a new example on the Z-test for the difference between two proportions. The Sure Value Convenience Stores, CardioGood Fitness, More Descriptive Choices Follow-up, and Clear Mountain State Surveys cases are included. There is a new online section on Effect Size.

Chapter 11—The chapter has a new Using Statistics scenario that relates to a mobile electronics merchandiser that replaces the Perfect Parachutes scenario. This chapter includes the Sure Value Convenience Stores, CardioGood Fitness, More Descriptive Choices Follow-up, and Clear Mountain State Surveys cases. It now includes an online section on fixed effects, random effects, and mixed effects models.

Chapter 12—The chapter includes many new problems. This chapter includes the Sure Value Convenience Stores, CardioGood Fitness, More Descriptive Choices Follow-up, and Clear Mountain State Surveys cases. The McNemar test and the Chi-square test for a standard deviation or variance are now online sections.

Chapter 13—The Using Statistics scenario has been updated and changed, with new data used throughout the chapter. This chapter includes the Brynne Packaging case.

Chapter 14—The online section on influence analysis has been moved into the text.

Chapter 15—This chapter includes the Sure Value Convenience Stores, Craybill Instrumentation, and More Descriptive Choices Follow-up cases.

Chapter 16—This chapter includes new data involving movie attendance in Section 16.3 and updated data for The Coca-Cola Company in Sections 16.4 through 16.6 and Wal-Mart Stores, Inc., in Section 16.7. In addition, most of the problems are new or updated.

Chapter 17—This is the new business analytics chapter already discussed in *Innovations* on page xxiv. This chapter has been designed so that the descriptive methods or any of the predictive analytics methods can be taught separately and apart from the rest of the chapter should time not permit coverage of the entire chapter.

Chapter 18—This chapter now includes some new problems.

Chapter 19—The “Statistical Applications in Quality Management” chapter has been renumbered as Chapter 19 and moved online, where it is available for download as explained in Appendix C.

Chapter 20—The “Decision Making” chapter has been renumbered as Chapter 20 and remains available for download as explained in Appendix C.

About Our Educational Philosophy

In *Our Starting Point* at the beginning of this preface, we stated that we are guided by these key learning principles:

- Help students see the relevance of statistics to their own careers by providing examples drawn from the functional areas in which they may be specializing.
- Emphasize interpretation of statistical results over mathematical computation.
- Give students ample practice in understanding how to apply statistics to business.
- Familiarize students with how to use statistical software to assist business decision making.
- Provide clear instructions to students for using statistical applications.

The following further explains these principles:

1. **Help students see the relevance of statistics to their own careers by providing examples drawn from the functional areas in which they may be specializing.** Students need a frame of reference when learning statistics, especially when statistics is not their major. That frame of reference for business students should be the functional areas of business, such as accounting, finance, information systems, management, and marketing. Each statistics topic needs to be presented in an applied context related to at least one of these functional areas. The focus in teaching each topic should be on its application in business, the interpretation of results, the evaluation of the assumptions, and the discussion of what should be done if the assumptions are violated.
2. **Emphasize interpretation of statistical results over mathematical computation.** Introductory business statistics courses should recognize the growing need to *interpret* statistical results that computerized processes create. This makes the interpretation of results more important than knowing how to execute the tedious hand calculations required to produce them.
3. **Give students ample practice in understanding how to apply statistics to business.** Both classroom examples and homework exercises should involve actual or realistic data as much as possible. Students should work with data sets, both small and large, and be encouraged to look beyond the statistical analysis of data to the interpretation of results in a managerial context.
4. **Familiarize students with how to use statistical software to assist business decision making.** Introductory business statistics courses should recognize that programs with statistical functions are commonly found on a business decision maker's desktop computer. Integrating statistical software into all aspects of an introductory statistics course allows the course to focus on interpretation of results instead of computations (see point 2).
5. **Provide clear instructions to students for using statistical applications.** Books should explain clearly how to use programs such as Microsoft Excel and Minitab with the study of statistics, without having those instructions dominate the book or distract from the learning of statistical concepts.

Student Resources

Student Solutions Manual, by Professor Pin Tian Ng of Northern Arizona University and accuracy checked by Annie Puciloski, provides detailed solutions to virtually all the even-numbered exercises and worked-out solutions to the self-test problems (ISBN-10: 0-321-92670-6; ISBN-13: 978-0-321-92670-8).

Online resources—The complete set of online resources are discussed fully in Appendix C, which also explains how to download these resources. These resources include the **Excel and Minitab Data Files** that contain the data used in chapter examples or named in problems and end-of-chapter cases; the **Excel Guide Workbooks** that contain templates or model solutions for applying Excel to a particular statistical method; the **Digital Cases PDF files** that support the end-of-chapter Digital Cases; the **Visual Explorations Workbooks** that interactively demonstrate various key statistical concepts; and the **PHStat** add-in that simplifies the use of Microsoft Windows or OS X Microsoft Excel with this book, as explained in Section EG.1.

The online resources also include the **Chapter Short Takes** and **Online Topic Sections** that expand and extend the discussion of statistical concepts worksheet-based solutions as well as the full text of two additional chapters, “Statistical Applications in Quality Management” and “Decision Making.”

Instructor Resources

The following supplements are among the resources available to adopting instructors at the Instructor’s Resource Center, located at www.pearsonhighered.com/irc.

- **Instructor’s Solutions Manual**, by Professor Pin Tian Ng of Northern Arizona University and accuracy checked by Annie Puciloski, includes solutions for end-of-section and end-of-chapter problems, answers to case questions, where applicable, and teaching tips for each chapter.
- **Lecture PowerPoint Presentations**, by Professor Patrick Schur of Miami University and accuracy checked by David Levine and Kathryn Szabat, are available for each chapter. The PowerPoint slides provide an instructor with individual lecture outlines to accompany the text. The slides include many of the figures and tables from the text. Instructors can use these lecture notes as is or can easily modify the notes to reflect specific presentation needs.
- **Test Bank**, by Professor Pin Tian Ng of Northern Arizona University, contains true/false, multiple-choice, fill-in, and problem-solving questions based on the definitions, concepts, and ideas developed in each chapter of the text.
- **TestGen® (www.pearsoned.com/testgen)** enables instructors to build, edit, print, and administer tests using a computerized bank of questions developed to cover all the objectives of the text. TestGen is algorithmically based, allowing instructors to create multiple but equivalent versions of the same question or test with the click of a button. Instructors can also modify test bank questions or add new questions. The software and test bank are available for download from Pearson Education’s online catalog.

MathXL®

MathXL® for Statistics Online Course (access code required) MathXL® is the homework and assessment engine that runs MyStatLab. (MyStatLab is MathXL plus a learning management system.)

With *MathXL for Statistics*, instructors can:

- Create, edit, and assign online homework and tests using algorithmically generated exercises correlated at the objective level to the textbook.
- Create and assign their own online exercises and import TestGen tests for added flexibility.
- Maintain records of all student work, tracked in MathXL’s online gradebook.

With *MathXL for Statistics*, students can:

- Take chapter tests in MathXL and receive personalized study plans and/or personalized homework assignments based on their test results.
- Use the study plan and/or the homework to link directly to tutorial exercises for the objectives they need to study.
- Access supplemental animations directly from selected exercises.
- Knowing that students often use external statistical software, we make it easy to copy our data sets, both from the eText and the MyStatLab questions, into StatCrunch™, Microsoft Excel, Minitab, and a variety of other software packages.

MathXL for Statistics is available to qualified adopters. For more information, visit www.mathxl.com or contact your Pearson representative.

MyStatLab™

MyStatLab™ Online Course (access code required) MyStatLab from Pearson is the world’s leading online resource for statistics learning, integrating interactive homework, assessment, and media in a flexible, easy to use format. MyStatLab is a course management systems that delivers **proven results** in helping individual students succeed.

- MyStatLab can be successfully implemented in any environment—lab-based, hybrid, fully online, traditional—and demonstrates the quantifiable difference that integrated usage has on student retention, subsequent success, and overall achievement.
- MyStatLab’s comprehensive online gradebook automatically tracks students’ results on tests, quizzes, and homework and in the study plan. Instructors can use the gradebook to provide positive feedback or intervene if students have trouble. Gradebook data can be easily exported to a variety of spreadsheet programs, such as Microsoft Excel. You can determine which points of data you want to export and then analyze the results to determine success.

MyStatLab provides **engaging experiences** that personalize, stimulate, and measure learning for each student. In addition to the resources below, each course includes a full interactive online version of the accompanying textbook.

- **Tutorial Exercises with Multimedia Learning Aids:** The homework and practice exercises in MyStatLab align with the exercises in the textbook, and they regenerate algorithmically to give students unlimited opportunity for practice and mastery. Exercises offer immediate helpful feedback, guided solutions, sample problems, animations, videos, and eText clips for extra help at the point of use.
- **MyStatLab Accessibility:** MyStatLab is compatible with the JAWS 12/13 screen reader and enables multiple-choice and free-response problem types to be read and interacted with via keyboard controls and math notation input.
- **StatTalk Videos:** Fun-loving statistician Andrew Vickers takes to the streets of Brooklyn, NY to demonstrate important statistical concepts through interesting stories and real-life events. This series of 24 fun and engaging videos will help students actually understand statistical concepts. Available with an instructor’s user guide and assessment questions.
- **Business Insight Videos:** Ten engaging videos show managers at top companies using statistics in their everyday work. Assignable question encourage debate and discussion.
- **Additional Question Libraries:** In addition to algorithmically regenerated questions that are aligned with your textbook, the MyStatLab courses come with two additional question libraries: **450 Getting Ready for Statistics** covers the developmental math topics students need for the course. These can be assigned as a prerequisite to other assignments, if desired. **1000 Conceptual Question Library** requires students to apply their statistical understanding.
- **Integration of Statistical Software:** We make it easy to copy our data sets, both from the eText and the MyStatLab questions, into software such as StatCrunch, Minitab, Excel, and more. Students have access to a variety of support tools—Technology Tutorial Videos, Technology Study Cards, and Technology Manuals for select titles—to learn how to effectively use statistical software.
- **Expert Tutoring:** Although many students describe the whole of MyStatLab as “like having your own personal tutor,” students also have access to live tutoring from Pearson, from qualified statistics instructors.
- **StatCrunch®:** MyStatLab integrates the web-based statistical software StatCrunch within the online assessment platform so that students can easily analyze data sets from exercises and the text. In addition, MyStatLab includes access to www.statcrunch.com, a website where users can access tens of thousands of shared data sets, conduct online surveys, perform complex analyses using the powerful statistical software, and generate compelling reports.

And, MyStatLab comes from an **experienced partner** with educational expertise and an eye on the future.

- Knowing that you are using a Pearson product means knowing that you are using quality content. That means that our eTexts are accurate and our assessment tools work. It means we are committed to making MyMathLab as accessible as possible.
- Whether you are just getting started with MyStatLab, or have a question along the way, we’re here to help you learn about our technologies and how to incorporate them into your course.

To learn more about how MyStatLab combines proven learning applications with powerful assessment, visit www.mystatlab.com or contact your Pearson representative.

StatCrunch® is powerful web-based statistical software that allows users to perform complex analyses, share data sets, and generate compelling reports of their data. The vibrant online community offers tens of thousands shared data sets for students to analyze.

Full access to StatCrunch is available with a MyStatLab access kit, and StatCrunch is available by itself to qualified adopters. StatCrunch is now compatible with most mobile devices. To access, visit www.statcrunch.com/mobile from the browser on your smartphone or tablet. For more information, visit our website at www.statcrunch.com, or contact your Pearson representative.

Acknowledgments

We are extremely grateful to the RAND Corporation and the American Society for Testing and Materials for their kind permission to publish various tables in Appendix E, and to the American Statistical Association for its permission to publish diagrams from the *American Statistician*.

A Note of Thanks

We thank Mohammad Ahmadi, University of Tennessee Chattanooga; Shubbo Bandyopadhyay, University of Florida; William Borders, Troy University; Steven Garren, James Madison University; Jersy Kamburowski, University of Toledo; M. B. Khan, California State University Long Beach; Hui Min Li, West Chester University; Nelson Modeste, Tennessee State University; Chris Morgan, Purdue University; Patricia Mullins, University of Wisconsin; Yvonne Sandoval, University of Arizona; and Yan Yu, University of Cincinnati for their comments, which have made this a better book. We also appreciate and acknowledge the assistance of Jian Cao and Curt Hinrichs of the SAS Institute in helping us prepare some of the contents of the new Chapter 17.

Creating a new edition of a textbook is a team effort, and we would like to thank our Pearson Education editorial, marketing, and production teammates: Sonia Ashraf, Dana Bettez, Kathleen DeChavez, Erin Lane, Deirdre Lynch, Kathy Manley, Christine Stavrou, Marianne Stepanian, and Joe Vetere. We also thank our statistical reader and accuracy checker Annie Puciloski for her diligence in checking our work and Nancy Kincade for overseeing and managing these efforts on behalf of PreMediaGlobal.

Finally, we would like to thank our families for their patience, understanding, love, and assistance in making this book a reality. It is to them that we dedicate this book.

Concluding Remarks

Please email us at authors@davidlevinestatistics.com if you have a question or require clarification about something discussed in this book. We also invite you to communicate any suggestions you may have for a future edition of this book. And while we have strived to make this book both pedagogically sound and error-free, we encourage you to contact us if you discover an error. When contacting us electronically, please include "BBS edition 13" in the subject line of your message.

You can also visit davidlevinestatistics.com, where you will find an email contact form and links to additional information about this book. For technical assistance using Microsoft Excel or any of the Excel add-ins that you can use with this book including PHStat, review Appendices D and G and follow the technical support links discussed in Appendix Section G.1, if necessary.

*Mark L. Berenson
David M. Levine
Kathryn A. Szabat*

This page intentionally left blank

GETTING STARTED

CONTENTS

- GS.1 Statistics: A Way of Thinking
- GS.2 Data: What Is It?
- GS.3 Business Analytics: The Changing Face of Statistics
“Big Data”
Statistics: An Important Part of Your Business Education

How to Use This Book

- GS.4 Software and Statistics

EXCEL GUIDE

- EG.1 Getting Started with Microsoft Excel
- EG.2 Entering Data
- EG.3 Opening and Saving Workbooks
- EG.4 Creating and Copying Worksheets
- EG.5 Printing Worksheets

MINITAB GUIDE

- MG.1 Getting Started with Minitab
- MG.2 Entering Data
- MG.3 Opening and Saving Worksheets and Projects
- MG.4 Creating and Copying Worksheets
- MG.5 Printing Parts of a Project

OBJECTIVES

- That the volume of data that exists in the world makes learning about statistics critically important
- That statistics is a way of thinking that can help you make better decisions
- How the DCOVA framework for applying statistics can help you solve business problems
- What business analytics is and how these techniques represent an opportunity for you
- How to make best use of this book
- How to prepare for using Microsoft Excel or Minitab with this book

Important Things to Learn First

USING STATISTICS

“You Cannot Escape from Data”

Not so long ago, business students were unfamiliar with the word *data* and had little experience handling data. Today, every time you visit a search engine website or “ask” your mobile device a question, you are handling data. And if you “check in” to a location or indicate that you “like” something, you are *creating* data as well.

You accept as almost true the premises of stories in which characters collect “a lot of data” to uncover conspiracies, to foretell disasters, or to catch a criminal. You hear concerns about how the government or business might be able to “spy” on you in some ways or how large social media companies “mine” your personal data for profit.

You hear the word *data* everywhere and may even have a “data plan” for your smartphone. You know, in a general way, that data are facts about the world and that most data seem to be, ultimately, a set of numbers—that 49% of students recently polled dreaded taking a business statistics course, or that 50% of citizens believe the country is headed in the right direction, or that unemployment is down 3%, or that your best friend’s social media account has 835 friends and 202 recent posts.

You cannot escape from data in this digital world. What, then, should you do? You could try to ignore data and conduct business by relying on hunches or your “gut feelings.” However, if you only want to use gut feelings, then you probably shouldn’t be reading this book or taking business courses in the first place.

You could note that there is so much data in the world—or just in your own little part of the world—that you couldn’t possibly get a handle on it. You could accept other people’s data summaries and their conclusions without first reviewing the data yourself. That, of course, would expose yourself to fraudulent practices.

Or, you could do things the proper way and realize that you cannot escape learning the methods of statistics, the subject of this book ...



Angela Waye/Shutterstock

GS.1 Statistics: A Way of Thinking

Statistics is a way of thinking that can help you make better decisions. Statistics helps you solve problems that involve decisions that are based on data that have been collected. You may have had some statistics instruction in the past. If you ever created a chart to summarize data or calculated values such as averages to summarize data, you have used statistics. But there's even more to statistics than these commonly taught techniques, as the detailed table of contents shows.

Statistics is undergoing important changes today. There are new ways of visualizing data that either did not exist, were not practical to do, or were not widely known until recently. And, more and more, statistics today is being used to "listen" to what the data might be telling you (the subject of Chapter 17) rather than just being a way to use data to prove something you want to say.

If you associate statistics with doing a lot of mathematical calculations, you will quickly learn that business statistics uses software to perform the calculations for you (and, generally, the software calculates with more precision and efficiency than you could do manually). But while you do not need to be a good manual calculator to apply statistics, because statistics is a way of thinking, you do need to follow a framework, or plan, to minimize possible errors of thinking and analysis. The **DCOVA framework** is one such framework.

THE DCOVA FRAMEWORK

The DCOVA framework consists of the following tasks:

- **Define** the data that you want to study in order to solve a problem or meet an objective.
- **Collect** the data from appropriate sources.
- **Organize** the data collected by developing tables.
- **Visualize** the data collected by developing charts.
- **Analyze** the data collected to reach conclusions and present those results.

The DCOVA framework uses the five tasks **Define**, **Collect**, **Organize**, **Visualize**, and **Analyze** to help apply statistics to business decision making. Typically, you do the tasks in the order listed. You must always do the first two tasks to have meaningful outcomes, but, in practice, the order of the other three can change or appear inseparable. Certain ways of visualizing data help you to organize your data while performing preliminary analysis as well. In any case, when you apply statistics to decision making, you should be able to identify all five tasks, and you should verify that you have done the first two tasks before the other three.

Using the DCOVA framework helps you to apply statistics to these four broad categories of business activities:

- Summarize and visualize business data
- Reach conclusions from those data
- Make reliable forecasts about business activities
- Improve business processes

Throughout this book, and especially in the Using Statistics scenarios that begin the chapters, you will discover specific examples of how DCOVA helps you apply statistics. For example, in one chapter, you will learn how to demonstrate whether a marketing campaign has increased sales of a product, while in another you will learn how a television station can reduce unnecessary labor expenses.

GS.2 Data: What Is It?

Defining data in a general way as “facts about the world,” to quote the opening essay, can prove confusing as such facts could be singular, a value associated with something, or collective, a list of values associated with something. For example, “David Levine” is a singular fact, a coauthor of this book, whereas “Mark, David, and Kathy” is the *collective* list of authors of this book. Furthermore, if everything is data, how do you distinguish “David Levine” from “Basic Business Statistics,” two very different facts (coauthor and title) about this book. Statisticians avoid this confusion by using a more specific definition of data and by defining a second word, *variable*.

Data are “the values associated with a trait or property that help distinguish the occurrences of something.” For example, the names “David Levine” and “Kathryn Szabat” are data because they are both values that help distinguish one of the authors of this book from another. In this book, *data* is always plural to remind you that data are a collection, or set, of values. While one could say that a single value, such as “David Levine,” is a *datum*, the phrases *data point*, *observation*, *response*, and *single data value* are more typically encountered.

The trait or property of something that values (data) are associated with is what statisticians define as a **variable**. For example, you might define the variables “coauthor” and “title” if you were defining data about a set of textbooks.

Substituting the word *characteristic* for the phrase “trait or property” and using the phrase “an item or individual” instead of the vague word “something” produces the definitions of *variable* and *data* used in this book.

Student Tip

Business convention places the data, the *set* of values, for a variable in a column when using a worksheet or similar object. The Excel and Minitab data worksheets used in this book follow this convention. Because of this convention, people sometimes use the word *column* as a substitute for *variable*.

VARIABLE

A characteristic of an item or individual.

DATA

The set of individual values associated with a variable.

Think about characteristics that distinguish individuals in a human population. Name, height, weight, eye color, marital status, adjusted gross income, and place of residence are all characteristics of an individual. All of these traits are possible *variables* that describe people.

Defining a variable called author-name to be the first and last names of the authors of this text makes it clear that valid values would be “Mark Berenson,” “David Levine,” and “Kathryn Szabat” and not, say, “Berenson,” “Levine,” and “Szabat.” Be careful of cultural or other assumptions in definitions—for example, is “last name” a family name, as is common usage in North America, or an individual’s own unique name, as is common usage in most Asian countries?

Having defined *data* and *variable*, you can define the subject of this book, **statistics**.

STATISTICS

The methods that help transform data into useful information for decision makers.

Statistics allows you to determine whether your data represent information that could be used in making better decisions. Therefore, statistics helps you determine whether differences in the numbers are meaningful in a significant way or are due to chance. To illustrate, consider the following news reports about various data findings:

- “Acceptable Online Ad Length Before Seeing Free Content” (*USA Today*, February 16, 2012, p. 1B) A survey of 1,179 adults 18 and over reported that 54% thought that 15 seconds was an acceptable online ad length before seeing free content.

- “First Two Years of College Wasted?” (M. Marklein, *USA Today*, January 18, 2011, p. 3A) A survey of more than 3,000 full-time, traditional-age students found that the students spent 51% of their time on socializing, recreation, and other activities; 9% of their time attending class/lab; and 7% of their time studying.
- “Follow the Tweets” (H. Rui, A. Whinston, and E. Winkler, *The Wall Street Journal*, November 30, 2009, p. R4) In this study, the authors found that the number of times a specific product was mentioned in comments in the Twitter social messaging service could be used to make accurate predictions of sales trends for that product.

Without statistics, you cannot determine whether the “numbers” in these stories represent useful information. Without statistics, you cannot validate claims such as the claim that the number of tweets can be used to predict the sales of certain products. And without statistics, you cannot see patterns that large amounts of data sometimes reveal.

When talking about statistics, you use the term **descriptive statistics** to refer to methods that primarily help summarize and present data. Counting physical objects in a kindergarten class may have been the first time you used a *descriptive* method. You use the term **inferential statistics** to refer to methods that use data collected from a small group to reach conclusions about a larger group. If you had formal statistics instruction in a lower grade, you were probably mostly taught descriptive methods, the focus of the early chapters of this book, and you may be unfamiliar with many of the inferential methods discussed in later chapters.

GS.3 Business Analytics: The Changing Face of Statistics

The Using Statistics scenario that opens this chapter notes the increasing use of new statistical techniques that either did not exist, were not practical to do, or were not widely known in the past. Of all these new techniques, business analytics best reflects the changing face of statistics. These methods combine traditional statistical methods with methods from management science and information systems to form an interdisciplinary tool that supports fact-based management decision making. Business analytics enables you to

- Use statistical methods to analyze and explore data to uncover unforeseen relationships.
- Use management science methods to develop optimization models that impact an organization’s strategy, planning, and operations.
- Use information systems methods to collect and process data sets of all sizes, including very large data sets that would otherwise be hard to examine efficiently.

Business analytics allows you to interpret data, reach conclusions, and make decisions and, in doing that, it combines many of the tasks of the DCOVA framework into one integrated process. And because you apply business analytics in the context of *organizational* decision making and problem solving (see reference 7), successful application of business analytics requires an understanding of a business and its operations. Chapter 17 examines business analytics more closely, including its implications for the future.

“Big Data”

Relatively recent advances in information technology allow businesses to collect, process, and analyze very large volumes of data. Because the operational definition of “very large” can be partially dependent on the context of a business—what might be “very large” for a sole proprietorship might be commonplace and small for a multinational corporation—many use the term *big data*.

Big data is more of a fuzzy concept than a term with a precise operational definition, but it implies data that are being collected in huge volumes and at very fast rates (typically in real time) and data that arrive in a variety of forms, organized and unorganized. These attributes of “volume, velocity, and variety,” first identified in 2001 (see reference 5), make big data different from any of the data sets used in this book.

Big data increases the use of business analytics because the sheer size of these very large data sets makes preliminary exploration of the data using older techniques impractical to do. This effect is explored in Chapter 17.

Statistics: An Important Part of Your Business Education

As business analytics becomes increasingly important in business, and especially as the use of big data increases, statistics, an essential component of business analytics, becomes increasingly important to your business education. In the current data-driven environment of business, you need general analytical skills that allow you to manipulate data, interpret analytical results, and incorporate results in a variety of decision-making applications, such as accounting, finance, HR management, marketing, strategy/planning, and supply chain management.

The decisions you make will be increasingly based on data and not on gut or intuition supported by personal experience. Data-guided practice is proving to be successful; studies have shown an increase in productivity, innovation, and competition for organizations that embrace business analytics. The use of data and data analysis to drive business decisions cannot be ignored. Having a well-balanced mix of technical skills—such as statistics, modeling, and basic information technology skills—and managerial skills—such as business acumen, problem-solving skills, and communication skills—will best prepare you for today's, and tomorrow's, workplace (see reference 1).

If you thought that you could artificially separate statistics from other business subjects, take a statistics course, and then forget about statistics, you have overlooked the changing face of statistics. The changing face is the reason that Hal Varian, the chief economist at Google, Inc., noted as early as 2009, “the sexy job in the next 10 years will be statisticians. And I’m not kidding” (see references 8 and 9).

How to Use This Book

This book helps you develop the skills necessary to use the DCOVA framework to apply statistics to the four types of business activities listed on page 2. Chapter 1 discusses the **D**efine and **C**ollect tasks, the necessary starting point for all statistical activities. Chapters 2 and 3 explain the **O**rganize and **V**isualize tasks and present methods that summarize and visualize business data (the first activity listed in Section GS.1). Chapter 3 also presents statistics used in the **A**nalyze task. Chapters 4 through 12 discuss methods that use sample data to reach conclusions about populations (the second activity listed). Chapters 13 through 16 review methods to make reliable forecasts (the third activity). The online Chapter 19 introduces methods that you can use to improve business processes (the fourth activity) and the online Chapter 20 introduces decision-making methods. (As previously noted, Chapter 17 discusses business analytics.) Chapter 18 summarizes the methods of this book and provides you with a roadmap for analyzing data.

Each chapter begins with a Using Statistics scenario that puts you in a realistic business situation. You will face problems that the statistical concepts and methods introduced in the chapter will help solve. Later, near the end of the chapter, a Using Statistics Revisited section reviews how the statistical methods discussed in the chapter can be applied to help solve the problems you faced.

Each chapter ends with a variety of features that help you review what you have learned in the chapter. Summary, Key Equations, and Key Terms concisely present the important points of a chapter.

Checking Your Understanding tests your understanding of basic concepts, and Chapter Review Problems allow you to practice what you have learned.

Throughout this book, you will find Excel and Minitab solutions to example problems. You will also find many *Student Tips*, margin notes that help clarify and reinforce significant details about particular statistical concepts. Selected chapters include Visual Explorations features that allow you to interactively explore statistical concepts. And many chapters include a “Think About This” essay that explains important statistical concepts in further depth.

This book contains numerous case studies that give you an opportunity to enhance your analytic and communication skills. Appearing in most chapters is the continuing case study *Managing Ashland MultiComm Services* that details problems managers of a residential telecommunications provider face and a Digital Case, which asks you to sort through information in electronic documents and then apply your statistical knowledge to resolve a business problem or issue. Besides these two cases, you will find a number of other cases, including some that reoccur in several chapters, in this book.

Don't worry if your instructor does not cover every section of every chapter. Introductory business statistics courses vary in terms of scope, length, and number of college credits earned. Your functional area of study (accounting, management, finance, marketing, etc.) may also affect what you learn in class or what you are assigned to read in this book.

GS.4 Software and Statistics

You use software to assist you in applying statistical methods to business decision making. Microsoft Excel and Minitab are examples of applications that people use for statistics. Excel is the Microsoft Office data analysis application that evolved from earlier electronic spreadsheets used in accounting and financial applications. Minitab, a dedicated statistical application, or **statistical package**, was developed from the ground up to perform statistical analysis as accurately as possible. Versions of Minitab run on larger computer systems and can perform sophisticated analyses of large data sets.

Although you are probably more familiar with Excel than with Minitab, both programs share many similarities, starting with their shared use of **worksheets** (or spreadsheets) to store data for analysis. Worksheets are tabular arrangements of data, in which the intersections of rows and columns form **cells**, boxes into which you make entries. In Minitab, the data for each variable are placed in separate columns, and this is also the standard practice when using Excel. Generally, to perform a statistical analysis in either program, you select one or more columns of data and then apply the appropriate command.

Both Excel and Minitab allow you to save worksheets, programming information, and results as one file, called a **workbook** in Excel and a **project** in Minitab. In Excel, workbooks are collections of worksheets and chart sheets. You save a workbook when you save “an Excel file” (either as an **.xlsx** or **.xls** file). In Minitab, a project includes data worksheets, all the results shown in a *session window*, and all graphs created for the data. Unlike in Excel, in Minitab you can save individual worksheets (as **.mtw** worksheet files) as well as save the entire project (as an **.mpj** project file).

Excel and Minitab Guides

You can use either Excel or Minitab to learn and practice the statistical methods learned in this book. Immediately following each chapter are Excel and Minitab Guides. For this chapter, special guides explain how the guides have been designed to support your learning with this book. To prepare for using Excel or Minitab, review and complete the checklist in Table GS.1 below.

TABLE GS.1

Checklist for Preparing to Use Excel or Minitab with This Book

- Determine which program, Excel or Minitab, you will use with this book.
- Read and review the Excel or Minitab Guide for this chapter to verify your knowledge of required basic skills.
- Read Appendix C to learn about the online resources you need to make best use of this book. Appendix C includes a complete list of the data files that are used in the examples and problems found in this book. Names of data files appear in this distinctive type face—**Retirement Funds**—throughout this book.
- Download the online resources that you will need to use this book, using the instructions in Appendix C.
- Check for updates to the program that you plan to use with this book, using the Appendix Section D.1 instructions.
- If you plan to use Excel with PHStat, the Visual Explorations add-in workbooks, or the Analysis ToolPak and you maintain your own computer system, read the special instructions in Appendix D.
- Examine Appendix G to learn answers to frequently asked questions (FAQs).

In later chapters, these guides are keyed to the in-chapter section numbers and present detailed Excel and Minitab instructions for performing the statistical methods discussed in chapter sections. Table GS.2 presents the typographic conventions that the guides use to present computer operations. Excel Guides additionally identify the key Excel technique that is used for a statistical method and include instructions for using PHStat, the Pearson Education statistics add-in that simplifies the operation of Microsoft Excel.

TABLE GS.2

Computing Conventions
Used in This Book

Operation and Examples	Notes
Keyboard keys Enter Ctrl Shift	Names of keys are always the object of the verb <i>press</i> , as in “press Enter .”
Keystroke combinations Ctrl+C Ctrl+Shift+Enter Command+Enter	Keyboarding actions that require you to press more than one key at the same time. Ctrl+C means press C while holding down Ctrl . Ctrl+Shift+Enter means press Enter while holding down both Ctrl and Shift .
Click or select operations click OK select the first 2-D Bar gallery item	Mouse pointer actions that require you to single click an onscreen object. This book uses the verb <i>select</i> when the object is either a worksheet cell or an item in a gallery, menu, list, or Ribbon tab.
Menu or ribbon selection File → New Layout → Legend → None	A sequence of Ribbon or menu selections. File → New means first select the File tab and then select New from the list that appears.
Placeholder object <i>variable 1 cell range</i> <i>bins cell range</i>	An italicized boldfaced phrase is a placeholder for an object reference. In making entries, you enter the reference, e.g., A1:A10 , and not the placeholder.

The guides presume that you have knowledge of the basic computing skills listed in Table GS.3. If you have not mastered these skills, you can read the online pamphlet *Basic Computing Skills*. (Appendix C explains how you can download a copy of this and other online sections.)

TABLE GS.3

Basic Computing Skills

Basic Skill	Specifics
Identification of application window objects	Title bar, minimize/resize/close buttons, scroll bars, formula bar, workbook area, cell pointer, shortcut menu. For Excel only, panes and these Ribbon parts: tab, group, gallery, and launcher button
Knowledge of mouse operations	Click (also called select), check and clear, double-click, right-click, drag/drag-and-drop
Identification of dialog box objects	Command button, list box, drop-down list, edit box, option button, check box

REFERENCES

1. Advani, D. “Preparing Students for the Jobs of the Future.” *University Business* (2011), www.universitybusiness.com/article/preparing-students-jobs-future.
2. Davenport, T., and J. Harris. *Competing on Analytics: The New Science of Winning*. Boston: Harvard Business School Press, 2007.
3. Davenport, T., J. Harris, and R. Morison. *Analytics at Work*. Boston: Harvard Business School Press, 2010.
4. Keeling, K., and R. Pavur. “Statistical Accuracy of Spreadsheet Software.” *The American Statistician* 65 (2011): 265–273.
5. Laney, D. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Stamford, CT: META Group. February 6, 2001.
6. Levine, D., and D. Stephan. “Teaching Introductory Business Statistics Using the DCOVA Framework.” *Decision Sciences Journal of Innovative Education* 9 (September 2011): 393–398.
7. Liberatore, M., and W. Luo. “The Analytics Movement.” *Interfaces* 40 (2010): 313–324.
8. Varian, H. “For Today’s Graduate, Just One Word: Statistics.” *The New York Times*, August 6, 2009, www.nytimes.com/2009/08/06/technology/06stats.html.
9. Varian, H. “Hal Varian and the Sexy Profession.” *Significance*, March 2011.

KEY TERMS

big data 4

cells 6

data 3

DCOVA framework 2

descriptive statistics 4

inferential statistics 4

project 6

statistical package 6

statistics 3

template 8

variable 3

workbook 6

worksheet 6

EXCEL GUIDE

EG.1 GETTING STARTED with MICROSOFT EXCEL

You can use Excel to learn and apply the statistical methods discussed in this book and as an aid in solving end-of-section and end-of-chapter problems. How you use Excel is up to you (or perhaps your instructor), and the Excel Guides give you two complementary ways to use Excel.

If you are focused more on getting results as quickly as possible, consider using PHStat. PHStat, available for users of this book, is an example of an add-in, an application that extends the functionality of Microsoft Excel. The PHStat add-in simplifies the task of operating Excel while creating *real* Excel worksheets that use in-worksheet calculations. With PHStat, you can create worksheets that are identical to the ones featured in this book while minimizing the potential for making worksheet entry errors. In contrast, most other add-ins create results that are mostly text pasted into an empty worksheet.

For many topics, you may choose to use the *In-Depth Excel* instructions. These instructions use pre-constructed worksheets as models or **templates** for a statistical solution. You learn how to adapt these worksheets to construct your own solutions. Many of these sections feature a specific *Excel Guide workbook* that contains worksheets that are *identical* to the worksheets that PHStat creates. Because both of these ways create the same results and the same worksheets, you can use a combination of both ways as you read through this book.

The In-Depth Excel instructions and the Excel Guide workbooks work best with the latest versions of Microsoft Excel, including Excel 2010 and Excel 2013 (Microsoft Windows), Excel 2011 (OS X), and Office 365. Where incompatibilities arise with versions older than Excel 2010, the incompatibilities are noted and alternative worksheets are provided for use. (Excel Guides also contain instructions for using the Analysis ToolPak add-in that is included with some Microsoft Excel versions, when appropriate.)

You will want to master the Table EG.A basic skills before you begin using Excel to understand statistical concepts and solve problems. If you plan to use the *In-Depth Excel* instructions, you will also need to master the skills listed in the second half of the table. While you do not necessarily need these skills if you plan to use PHStat, knowing them will be useful if you expect to customize the Excel worksheets that PHStat creates or expect to be using Excel beyond the course that uses this book.

TABLE EG.A

Skills Set for Using Microsoft Excel with This Book

Basic Microsoft Office Skill	Specifics
Excel data entry	Organizing worksheet data in columns, entering numerical and categorical data
File operations	Open, save, print
Worksheet operations	Create, copy
In-Depth Excel Skill	Specifics
Formula skills	Concept of a formula, cell references, absolute and relative cell references, how to enter a formula, how to enter an array formula
Workbook presentation	How to apply format changes that affect the display of worksheet cell contents
Chart formatting correction	How to correct the formatting of charts that Excel improperly creates
Discrete histogram creation	How to create a properly formatted histogram for a discrete probability distribution

This guide reviews the basic Microsoft Office skills and Appendix B teaches you the *In-Depth Excel* skills. If you start by studying Sections B.1 through B.4 of that appendix, you will have the skills you need to make effective use of the *In-Depth Excel* instructions when you first encounter them in Chapter 1. (You can read other sections in Appendix B as needed.)

EG.2 ENTERING DATA

As noted in Section GS.4, you enter data into the rows and columns of a worksheet. By convention, and the style used in this book, when you enter data for a set of variables, you enter the name of each variable into the cells of the first row, beginning with column A. Then you enter the data for the variable in the subsequent rows to create a DATA worksheet similar to the one shown in Figure EG.1.

FIGURE EG.1

An example of a DATA worksheet

	A	B	C	D	E	F	G	H	I	J	K	L
1	Fund Number	Market Cap	Type	Assets	Turnover Ratio	Beta	SD	Risk	1YrReturn%	3YrReturn%	5YrReturn%	10YrReturn%
2	RF001	Large	Growth	309.90	12.21	1.15	18.72	Low	28.99	24.26	11.06	8.97
3	RF002	Large	Growth	23.30	0.00	2.19	35.72	High	33.40	22.72	-4.89	0.02
4	RF003	Large	Growth	141.50	147.00	2.24	36.69	High	33.98	21.91	1.53	12.55
5	RF004	Large	Growth	118.50	5.00	2.24	36.63	High	33.78	21.89	1.57	12.69
6	RF005	Large	Growth	575.30	121.00	0.89	14.56	Low	21.62	16.47	9.40	10.30

Student Tip

Most of the Excel data workbooks that you can download and use with this book (see Appendix C) contain a DATA worksheet that follows the rules of this section. You can use any of those worksheets as an additional model for data entry.

To enter data in a specific cell, either use the cursor keys to move the cell pointer to the cell or use your mouse to select the cell directly. As you type, what you type appears in the formula bar. Complete your data entry by pressing **Tab** or **Enter** or by clicking the checkmark button in the formula bar.

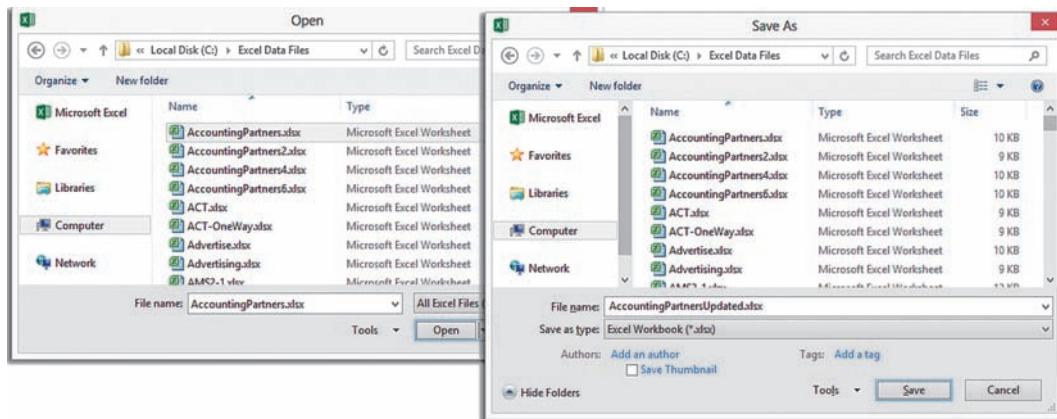
When you enter data, never skip any rows in a column, and as a general rule, also avoid skipping any columns. Also try to avoid using numbers as row 1 variable headings; if you cannot avoid their use, precede such headings with apostrophes. Pay attention to any special instructions that occur throughout the book for the order of the entry of your data. For some statistical methods, entering your data in an order that Excel does not expect will lead to incorrect results.

EG.3 OPENING and SAVING WORKBOOKS

You open and save a workbook by first selecting the folder that stores the workbook and then specifying the file name of the workbook. In most Excel versions, select **File → Open** to open a workbook file and **File → Save As** to save a workbook. (In Excel 2007, select **Office Button → Open** to open a workbook file and **Office Button → Save As** to save a workbook.) **Open** and **Save As** display nearly identical dialog boxes that vary only slightly among the different Excel versions. Figure EG.2 shows the Excel 2013 Open and Save As dialog boxes. To see these dialog boxes in Excel 2013, double-click **Computer** in the Open or Save As panels, a step that other Excel versions do not require.

FIGURE EG.2

Excel 2013 Open and Save As dialog boxes



You select the storage folder by using the drop-down list at the top of either of these dialog boxes. You enter, or select from the list box, a file name for the workbook in the **File name** box. You click **Open** or **Save** to complete the task. Sometimes when saving files, you may want to change the file type before you click **Save**.

In Microsoft Windows Excel versions, to save your workbook in the format used by versions older than Excel 2007, select **Excel 97-2003 Workbook (*.xls)** from the **Save as type** drop-down list before you click **Save**.

To save data in a form that can be opened by programs that cannot open Excel workbooks, you might select either **Text (Tab delimited) (*.txt)** or **CSV (Comma delimited) (*.csv)** as the save type. In OS X Excel versions, the equivalent selections are to select **Excel 97–2004 Workbook (.xls)**, **Tab Delimited Text (.txt)**, or **Windows Comma Separated (.csv)** from the **Format** drop-down list before you click **Save**.

When you want to open a file and cannot find its name in the list box, double-check that the current folder being searched is the proper folder. If it is, change the file type to **All Files (*.*)** (**All Files** in OS X Excel) to see all files in the current folder. This technique can help you discover inadvertent misspellings or missing file extensions that otherwise prevent the file from being displayed.

Although all versions of Microsoft Excel include a **Save** command, you should avoid this choice until you gain experience. Using **Save** makes it too easy to inadvertently overwrite your work. Also, you cannot use the **Save** command for any open workbook that Excel has marked as read-only. (Use **Save As** to save such workbooks.)

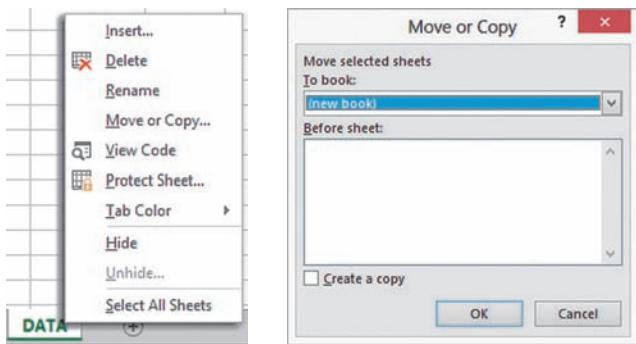
EG.4 CREATING and COPYING WORKSHEETS

You create new worksheets by either creating a new workbook or by inserting a new worksheet in an open workbook. In Microsoft Windows Excel versions, select **File** → **New** (**Office Button** → **New** in Excel 2007) and in the pane that appears, double-click the Blank workbook icon. In OS X Excel 2011, select **File** → **New Workbook**.

New workbooks are created with a fixed number of worksheets. To delete extra worksheets or insert more sheets, right-click a sheet tab and click either **Delete** or **Insert** (see Figure EG.3). By default, Excel names a worksheet serially, in the form Sheet1, Sheet2, and so on. You should change these names to better reflect the content of your worksheets. To rename a worksheet, double-click the sheet tab of the worksheet, type the new name, and press **Enter**.

FIGURE EG.3

Worksheet tab shortcut menu (left) and the Move or Copy dialog box (right)



You can also make a copy of a worksheet or move a worksheet to another position in the same workbook or to a second workbook. Right-click the sheet tab and select **Move or Copy** from the shortcut menu that appears. In the **To book** drop-down list of the Move or Copy dialog box (see Figure EG.3), first select **(new book)** (or the name of the pre-existing target workbook), check **Create a copy**, and then click **OK**.

EG.5 PRINTING WORKSHEETS

Student Tip

Although every version of Excel offers the (print) Entire workbook choice, you get the best results if you print each worksheet separately when you need to print more than one worksheet (or chart sheet).

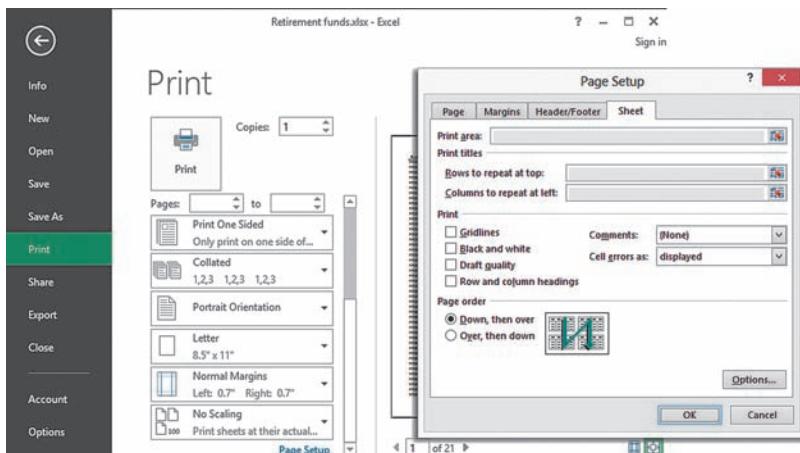
FIGURE EG.4

Excel 2013 Print Preview and Page Setup (inset) dialog boxes

To print a worksheet (or a chart sheet), click its sheet tab to open to the sheet. Then, in all Excel versions except Excel 2007, select **File** → **Print**. If the print preview (partially obscured in Figure EG.4) is acceptable to you, click the **Print** button. To return to the worksheet, press **Esc** (Excel 2013), click **File** (Excel 2010), or **Cancel** (OS X Excel 2011).

If necessary, you can adjust print formatting while in print preview by clicking **Page Setup** to display the Page Setup dialog box (see Figure EG.4 inset). For example, to print your worksheet with gridlines and numbered row and lettered column headings (similar to the appearance of the worksheet onscreen), click the **Sheet** tab in the Page Setup dialog box, check **Gridlines** and **Row and column headings**, and click **OK**.

In Excel 2007, printing requires additional mouse clicks. First click **Office Button** and then move the mouse pointer over (but do not click) **Print**. In the Preview and Print gallery, click **Print Preview**. If the preview contains errors or displays the worksheet in an undesirable manner, click **Close Print Preview**, make the necessary changes, and reselect the **Print Preview**. After completing all corrections and adjustments, click **Print** in the Print Preview window to display the Print dialog box. Select the printer to be used from the **Name** drop-down list, click **All** and **Active sheet(s)**, adjust the **Number of copies**, and click **OK**.



MINITAB GUIDE

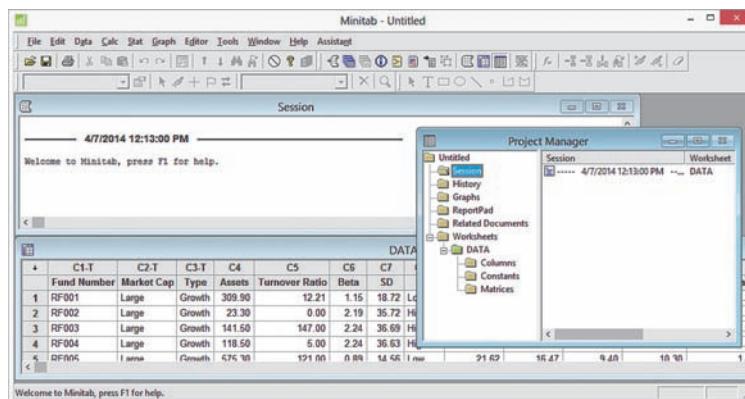
MG.1 GETTING STARTED with MINITAB

You can use Minitab to learn and apply the statistical methods discussed in this book and as an aid in solving end-of-section and end-of-chapter problems. As explained in Section GS.4, in Minitab you create and use project files that contain worksheets and other components.

When you start Minitab, you typically see a new project that contains only the session window and one worksheet window. These windows appear *inside* the Minitab window, and sometimes you may need to adjust the size of the Minitab window to see the entire window of a project component. In Figure MG.1, the **Project Manager** window that summarizes the content of the current project overlaps the session and DATA worksheet windows. You can arrange or hide these windows as you see fit. To view a particular window that may be obscured or hidden, select **Window** from the Minitab menu bar, and then select the name of the window you want to make visible.

FIGURE MG.1

Minitab main worksheet with overlapping session, worksheet, and Project Manager windows



MG.2 ENTERING DATA

Minitab uses the standard business convention, expecting data for a variable to be entered into a column. In this book, data are entered in columns, left to right, starting with the first column. Column names take the form C_n , such that the first column is named C_1 , the second column is C_2 , and the tenth column is C_{10} . Column names appear in the top border of a Minitab worksheet. Columns that contain non-numerical data have names that include “-T” (C_{2-T} and C_{3-T} in Figure MG.1). Columns that contain data that Minitab interprets as either dates or times have names that include “-D” (not seen in Figure MG.1).

When entering data, you use the first, unnumbered and shaded row to enter variable names. You can then refer to the column by that name or its C_n name in Minitab procedures. If a variable name contains spaces or other special characters, such as **Market Cap**, Minitab will display that name in dialog boxes using a pair of single quotation marks ('**Market Cap**'). You must include those quotation marks any time you enter such a variable name in a dialog box.

To enter or edit data in a specific cell, either use the cursor keys to move the cell pointer to the cell or use your mouse to select the cell directly. Never skip any rows in a column, and as a general rule, also avoid skipping any columns. Minitab interprets a skipped row as holding a “missing” value in the sense discussed in Section 1.3.

MG.3 OPENING and SAVING WORKSHEETS and PROJECTS

You open and save Minitab worksheet or project files by first selecting the folder that stores a workbook and then specifying the file name of the workbook. To open a worksheet, select **File → Open Worksheet**. To open a project, select **File → Open Project**. To save a worksheet, select **File → Save Current Worksheet As**. To save a project, select **File → Save Project As**. Both pairs of open and save commands display nearly identical dialog boxes. Figure MG.2 shows the Minitab 16 Open Worksheet and Save Current Worksheet As dialog boxes.

Inside the Open or Save dialog boxes, you select the storage folder by using the drop-down list at the top of either dialog box. You enter or select from the list box a file name for the workbook in the **File name** box. You click **Open** or **Save** to complete the task. Sometimes when saving files, you might want to change the file type before you click **Save**. If you want to save your data as an Excel worksheet, select **Excel** from the **Save as type** drop-down list before you click **Save**. If you want to save data in a form that

Student Tip

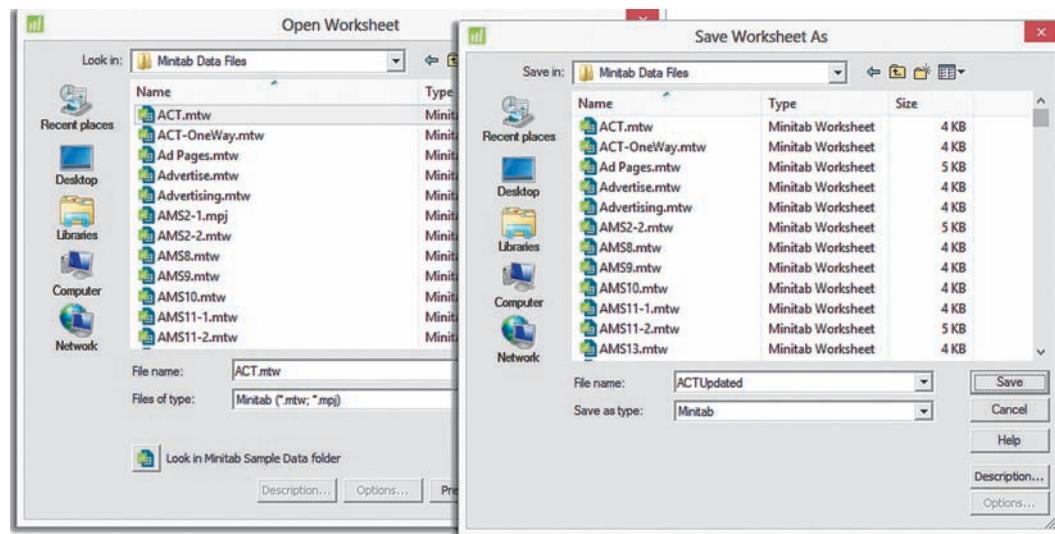
The Minitab worksheets that you can download and use with this book (see Appendix C) follow the rules of this section. You can use any of those worksheets as an additional model for data entry.

Student Tip

When you save a project, you can click **Options** in the Save Project As dialog box to open the Save Project - Options dialog box to selectively save parts of a project.

FIGURE MG.2

Minitab 16 Open Worksheet and Save Current Worksheet As dialog boxes



can be opened by programs that cannot open Excel workbooks, you might select one of the **Text** or **CSV** choices as the **Save as type** type.

When you want to open a file and cannot find its name in the list box, double-check that the current **Look in** folder is the folder you intend. If it is, change the file type to **All (*.*)** to see all files in the current folder. This technique can help you discover inadvertent misspellings or missing file extensions that otherwise prevent the file from being displayed.

Although Minitab includes **Save Current Worksheet** and a **Save Project** commands (commands without the “As”), you should avoid this choice until you gain experience. Using Save makes it too easy to inadvertently overwrite your work. Also, you cannot use the Save command for any open workbook that Minitab has marked as read-only. (Use Save As to save such workbooks.)

Individual graphs and a project’s session window can also be opened and saved separately in Minitab, although these operations are never used in this book.

MG.4 CREATING and COPYING WORKSHEETS

You create new worksheets by either creating a new project or by inserting a new worksheet in an open project. To create a new project, select **File → New** and in the New dialog box, click **Minitab Project** and then click **OK**. To insert a new worksheet, also select **File → New** but in the New dialog box click **Minitab Worksheet** and then click **OK**.

A new project is created with one new worksheet. To insert another worksheet, select **File → New** and in the New dialog box click **Minitab Worksheet** and then click **OK**. You can also insert a copy of a worksheet from another project into the current project. Select **File → Open Worksheet** and select the *project* that contains the worksheet to be copied. Selecting a project (and not a worksheet) causes an additional dialog box to be displayed, in which you can specify which worksheets of that second project are to be copied and inserted into the current project.

By default, Minitab names a worksheet serially in the form Worksheet1, Worksheet2, and so on. You should change these names to better reflect the content of your worksheets. To rename a worksheet, open the **Project Manager** window (see Figure MG.1), right-click the worksheet name in the left pane, select **Rename** from the shortcut menu, type in the new name, and press **Enter**. You can also use the **Save Current Worksheet As** command discussed in Section MG.3, although this command also saves the worksheet as a separate file.

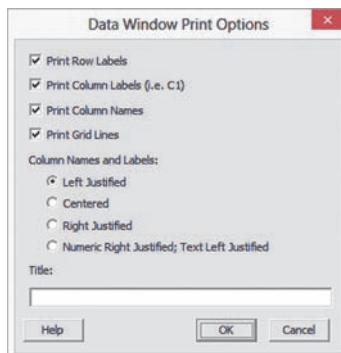
MG.5 PRINTING PARTS of a PROJECT

To print a worksheet, a graph, or the contents of a session, first select the window that corresponds to the object you want to print. Then select **File → Print object**, where *object* is either **Worksheet**, **Graph**, or **Session Window**, depending on which object you first selected.

If you are printing a graph or a session window, selecting the **Print** command displays the Print dialog box. The Print dialog box contains settings to select the printer to be used, what pages to print, and the number of copies to produce. If you need to change these settings, change them before clicking **OK** to create your printout.

If you are printing a worksheet, selecting **Print Worksheet** displays the Data Window Print Options dialog box (shown below). In this dialog box, you specify the formatting options for your printout (the default selections should be fine), enter a **Title**, and click **OK**. Minitab then presents the Print dialog box discussed in the previous paragraph.

If you need to change the paper size or paper orientation of your printout, select **File → Print Setup** before you select the Print command, make the appropriate selections in the dialog box that appears, and click **OK**.



CHAPTER

1

Defining and Collecting Data

CONTENTS

- 1.1 Defining Data
- 1.2 Measurement Scales for Variables
- 1.3 Collecting Data
- 1.4 Types of Sampling Methods
- 1.5 Types of Survey Errors

THINK ABOUT THIS: New Media Surveys/Old Sampling Problems

USING STATISTICS: Beginning of the End ... Revisited

CHAPTER 1 EXCEL GUIDE

CHAPTER 1 MINITAB GUIDE

OBJECTIVES

- To understand the types of variables used in statistics
- To know the different measurement scales
- To know how to collect data
- To know the different ways to collect a sample
- To understand the types of survey errors

USING STATISTICS

Beginning of the End ... Or the End of the Beginning?

The past few years have been challenging for Good Tunes & More (GT&M), a business that traces its roots to Good Tunes, a store that exclusively sold music CDs and vinyl records.

GT&M first broadened its merchandise to include home entertainment and computer systems (the “More”), and then undertook an expansion to take advantage of prime locations left empty by bankrupt former competitors. Today, GT&M finds itself at a crossroads. Hoped-for increases in revenues that have failed to occur and declining profit margins due to the competitive pressures of online sellers have led management to reconsider the future of the business.

While some investors in the business have argued for an orderly retreat, closing stores and limiting the variety of merchandise, GT&M CEO Emma Levia has decided in a time of uncertainty to “double down” and expand the business by purchasing Whitney Wireless, a successful three-store chain that sells smartphones and other mobile devices.

Levia foresees creating a brand new “A-to-Z” electronics retailer but first must establish a fair and reasonable price for the privately held Whitney Wireless. To do so, she has asked a group of analysts to identify, define, and collect the data that would be helpful in setting a price for the wireless business. As part of that group, you quickly realize that you need the data that would help to verify the contents of the wireless company’s basic financial statements.

You focus on data associated with the company’s profit and loss statement and quickly realize the need for sales and expense-related variables. You begin to think about what the data for such variables would look like and how to collect those data. You realize that you are starting to apply the DCOVA framework to the objective of helping Levia acquire Whitney Wireless.



Tyler Olson/Shutterstock

Defining a business objective is only the beginning of the process of business decision making. In the GT&M scenario, the objective is to establish a fair and reasonable price for the company to be acquired. Establishing a business objective always precedes the application of statistics to business decision making. Business objectives can arise from any level of management and can be as varied as the following:

- A marketing analyst needs to assess the effectiveness of a new television advertisement.
- A pharmaceutical company needs to determine whether a new drug is more effective than those currently in use.
- An operations manager wants to improve a manufacturing or service process.
- An auditor wants to review the financial transactions of a company in order to determine whether the company is in compliance with generally accepted accounting principles.

Establishing an objective is the end of what some would label *problem* definition, the formal beginning of any business decision-making process. But establishing the objective also marks a beginning—of applying the DCOVA framework to the task at hand.

Recall from Section GS.1 that the DCOVA framework uses the five tasks **Define**, **Collect**, **Organize**, **Visualize**, and **Analyze** to help apply statistics to business decision making. Restated, using the definition of a variable on page 2, the DCOVA framework consists of these tasks:

- **Define** the *variables* that you want to study in order to solve a problem or meet an objective.
- **Collect** the data *for those variables* from appropriate sources.
- **Organize** the data collected by developing tables.
- **Visualize** the data collected by developing charts.
- **Analyze** the data collected to reach conclusions and present those results.

In this chapter, you will learn more about the **Define** and **Collect** tasks.

1.1 Defining Data

Defining the variables that you want to study in order to solve a problem or meet an objective, the **D** task in the DCOVA framework, involves more than just making a list of things to study. For each variable of interest that you identify you must supply an **operational definition**, a universally accepted meaning that is clear to all associated with an analysis. This definition should clearly identify the values of the variable necessary to ensure that collected data are acceptable and appropriate for analysis.

For example, in the “Beginning of the End ...” scenario, sales per year data might be of interest to you as you focus on the data associated with Whitney Wireless. But the variable yearly sales might be subject to miscommunication: Does the variable refer to sales per year for the entire chain or just to one individual store? Are the values of the variable defined as dollar sales or unit sales? Providing an operational definition assures that such miscommunication does not occur.

When defining categorical variables, even individual values for a variable may need to be defined. For example, in a famous example, researchers collecting demographic data asked persons to fill in a form, one line of which asked about sex. More than one person supplied the answer *yes* and not the *male* or *female* value that the researchers intended. (Perhaps this is the reason that such a variable is more typically today named *Gender*—gender’s operational definition is more self-apparent.)

Establishing the Variable Type

As part of defining a variable, you establish whether the variable is a *categorical* or *numerical* variable. **Categorical variables** (also known as **qualitative variables**) have values that can only be placed into categories such as yes and no. Gender (male or female) is a categorical

Student Tip

Providing operational definitions are important in other contexts such as when writing a book about business statistics. Go back to page 3 and discover the first operational definitions of this book, for the easily misconstrued words *variable* and *data*! Like those definitions, all important operational definitions in this book are highlighted in boxes similar to one found on page 3.

variable. So, too are “Do you have a Facebook profile?” (yes or no) and Student class designation (Freshman, Sophomore, Junior, or Senior). **Numerical variables** (also known as **quantitative variables**) have values that represent a counted or measured quantity.

Numerical variables are further identified as being either *discrete* or *continuous* variables. **Discrete variables** have numerical values that arise from a counting process. “Number of items purchased” is a discrete numerical variable because its values represent the count of the number of items purchased. **Continuous variables** have numerical values that arise from a measuring process. “The time spent waiting on a checkout line” is an example of a continuous numerical variable because its values can represent a measurement with a stopwatch.

Values of a continuous variable can take on any value within a continuum or an interval, depending on the precision of the measuring instrument. For example, the waiting time could be 1 minute, 1.1 minutes, 1.11 minutes, or 1.113 minutes, depending on the precision of the stopwatch used. (Theoretically, a perfectly precise measuring device would never generate two identical continuous values for the same variable. However, because no measuring device is so precise, identical continuous values can occur.)

At first glance, identifying the variable type may seem easy, but some variables that you might want to study could be either categorical or numerical, depending on how you define them. For example, “age” would seem to be an obvious numerical variable, but what if you are interested in comparing the buying habits of children, young adults, middle-aged persons, and retirement-age people? In that case, defining “age” as a categorical variable would make better sense. Again, this illustrates the earlier point that without operational definitions, variables are meaningless.

Asking questions about the variables you have identified for study can often be a great help in determining the type of variable you have. Table 1.1 illustrates the process.

TABLE 1.1

Identifying Types of Variables

Question	Responses	Data Type
Do you have a Facebook profile?	<input type="checkbox"/> Yes <input type="checkbox"/> No	Categorical
How many text messages have you sent in the past three days?	_____	Numerical (discrete)
How long did it take to download the update for your newest mobile app?	_____ seconds	Numerical (continuous)

1.2 Measurement Scales for Variables

Variables can be further identified by the level of measurement, or **measurement scale**. Statisticians use the terms *nominal scale* and *ordinal scale* to describe the values for a categorical variable and use the terms *interval scale* and *ratio scale* to describe the values for a numerical variable.

Nominal and Ordinal Scales

Values for a categorical variable are measured on a nominal scale or on an ordinal scale. A **nominal scale** (see Table 1.2) classifies data into distinct categories in which no ranking is implied. Examples of a nominal scaled variable are your favorite soft drink, your political party affiliation, and your gender. Nominal scaling is the weakest form of measurement because you cannot specify any ranking across the various categories.

An **ordinal scale** classifies values into distinct categories in which ranking is implied. For example, suppose that GT&M conducted a survey of customers who made a purchase and asked the question “How do you rate the overall service provided by Good Tunes & More during your most recent purchase?” to which the responses were “excellent,” “very good,” “fair,”

LEARN MORE

Read the SHORT TAKES for Chapter 1 to learn more about nominal and ordinal scales.

TABLE 1.2

Examples of Nominal Scales

Categorical Variable	Categories
Do you have a Facebook profile?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Type of investment	<input type="checkbox"/> Cash <input type="checkbox"/> Mutual funds <input type="checkbox"/> Other
Cellular provider	<input type="checkbox"/> AT&T <input type="checkbox"/> Sprint <input type="checkbox"/> Verizon <input type="checkbox"/> Other <input type="checkbox"/> None

and “poor.” The answers to this question represent an ordinal scaled variable because the responses “excellent,” “very good,” “fair,” and “poor” are ranked in order of satisfaction. Table 1.3 lists other examples of ordinal scaled variables.

TABLE 1.3

Examples of Ordinal Scales

Categorical Variable	Ordered Categories
Student class designation	Freshman Sophomore Junior Senior
Product satisfaction	Very unsatisfied Fairly unsatisfied Neutral Fairly satisfied Very satisfied
Faculty rank	Professor Associate Professor Assistant Professor Instructor
Standard & Poor’s investment grade ratings	AAA AA+ AA AA- A+A BBB
Course grade	A B C D F

Ordinal scaling is a stronger form of measurement than nominal scaling because an observed value classified into one category possesses more of a property than does an observed value classified into another category. However, ordinal scaling is still a relatively weak form of measurement because the scale does not account for the amount of the differences between the categories. The ordering implies only which category is “greater,” “better,” or “more preferred”—not by how much.

Interval and Ratio Scales

Values for a numerical variable are measured on an interval scale or a ratio scale. An **interval scale** (see Table 1.4) is an ordered scale in which the difference between measurements is a meaningful quantity but does not involve a true zero point. For example, a noontime temperature reading of 67° Fahrenheit is 2 degrees warmer than a noontime reading of 65°. In addition, the 2° Fahrenheit difference in the noontime temperature readings is the same as if the two noontime temperature readings were 74° and 76° Fahrenheit because the difference has the same meaning anywhere on the scale.

TABLE 1.4

Examples of Interval and Ratio Scales

Numerical Variable	Level of Measurement
Temperature (in degrees Celsius or Fahrenheit)	Interval
ACT or SAT standardized exam score	Interval
File download time (in seconds)	Ratio
Age (in years or days)	Ratio
Cost of a computer system (in U.S. dollars)	Ratio

A **ratio scale** is an ordered scale in which the difference between the measurements involves a true zero point, as in height, weight, age, or salary measurements. If GT&M conducted a survey and asked how much money you expected to spend on audio equipment in the next year, the responses to such a question would be an example of a ratio-scaled variable. A person who expects to spend \$1,000 on audio equipment expects to spend twice as much money as someone who expects to spend \$500. As another example, a person who weighs 240 pounds is twice as heavy as someone who weighs 120 pounds.

Temperature is a trickier case: Fahrenheit and Celsius scales are interval but not ratio scales; the “zero” value is arbitrary, not real. You cannot say that a noontime temperature reading of 4° Fahrenheit is twice as hot as 2° Fahrenheit. In contrast, a Kelvin temperature reading is ratio scaled. In this scale, the temperature 0° Kelvin means no molecular motion.

Data measured on an interval scale or on a ratio scale constitute the highest levels of measurement. They are stronger forms of measurement than an ordinal scale because you can determine not only which observed value is the largest but also by how much.

LEARN MORE

Read the SHORT TAKES for Chapter 1 to learn more about interval and ratio scales.

Problems for Sections 1.1 and 1.2

LEARNING THE BASICS

1.1 Four different beverages are sold at a fast-food restaurant: soft drinks, tea, coffee, and bottled water.

- Explain why the type of beverage sold is an example of a categorical variable.
- Explain why the type of beverage sold is an example of a nominal scaled variable.

1.2 U.S. businesses are listed by size: small, medium, and large. Explain why business size is an example of an ordinal scaled variable.

1.3 The time it takes to download a video from the Internet is measured.

- Explain why the download time is a continuous numerical variable.
- Explain why the download time is a ratio-scaled variable.

APPLYING THE CONCEPTS



1.4 For each of the following variables, determine whether the variable is categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous. In addition, determine the measurement scale.

- Number of cellphones in the household
- Monthly data usage (in MB)
- Number of text messages exchanged per month
- Voice usage per month (in minutes)
- Whether the cellphone is used for email

1.5 The following information is collected from students upon exiting the campus bookstore during the first week of classes.

- Amount of time spent shopping in the bookstore
- Number of textbooks purchased
- Academic major
- Gender

Classify each of these variables as categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous. In addition, determine the measurement scale for each of these variables.

1.6 For each of the following variables, determine whether the variable is categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous. In addition, determine the measurement scale for each variable.

- Name of Internet service provider
- Time, in hours, spent surfing the Internet per week
- Whether the individual uses a mobile phone to connect to the Internet
- Number of online purchases made in a month
- Where the individual uses social networks to find sought-after information

1.7 For each of the following variables, determine whether the variable is categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous. In addition, determine the measurement scale for each variable.

- Amount of money spent on clothing in the past month
- Favorite department store
- Most likely time period during which shopping for clothing takes place (weekday, weeknight, or weekend)
- Number of pairs of shoes owned

1.8 Suppose the following information is collected from Robert Keeler on his application for a home mortgage loan at the Metro County Savings and Loan Association.

- Monthly payments: \$2,227
- Number of jobs in past 10 years: 1
- Annual family income: \$96,000
- Marital status: Married

Classify each of the responses by type of data and measurement scale.

1.9 One of the variables most often included in surveys is income. Sometimes the question is phrased “What is your income (in thousands of dollars)?” In other surveys, the respondent is asked to “Select the circle corresponding to your income level” and is given a number of income ranges to choose from.

- In the first format, explain why income might be considered either discrete or continuous.
- Which of these two formats would you prefer to use if you were conducting a survey? Why?

1.10 If two students score a 90 on the same examination, what arguments could be used to show that the underlying variable—test score—is continuous?

1.11 The director of market research at a large department store chain wanted to conduct a survey throughout a metropolitan area

to determine the amount of time working women spend shopping for clothing in a typical month.

- a. Indicate the type of data the director might want to collect.
- b. Develop a first draft of the questionnaire needed in (a) by writing three categorical questions and three numerical questions that you feel would be appropriate for this survey.

1.3 Collecting Data

After defining the variables that you want to study, you can proceed with the data collection task. Collecting data is a critical task because if you collect data that are flawed by biases, ambiguities, or other types of errors, the results you will get from using such data with even the most sophisticated statistical methods will be suspect or in error. (For a famous example of flawed data collection leading to incorrect results, read the Think About This essay on page 26.)

Data collection consists of identifying data sources, deciding whether the data you collect will be from a population or a sample, cleaning your data, and sometimes recoding variables. The rest of this section explains these aspects of data collection.

Data Sources

You collect data from either primary or secondary data sources. You are using a **primary data source** if you collect your own data for analysis. You are using a **secondary data source** if the data for your analysis have been collected by someone else.

You collect data by using any of the following:

- Data distributed by an organization or individual
- The outcomes of a designed experiment
- The responses from a survey
- The results of conducting an observational study
- Data collected by ongoing business activities

Market research companies and trade associations distribute data pertaining to specific industries or markets. Investment services such as Mergent, Inc., provide business and financial data on publicly listed companies. Syndicated services sold by The Nielsen Company provide consumer research data to telecom and mobile media companies. Print and online media companies also distribute data that they may have collected themselves or may be republishing from other sources.

The outcomes of a designed experiment are a second data source. For example, a consumer goods company might conduct an experiment that compares the stain-removing abilities of several laundry detergents. Note that developing a proper experimental design is mostly beyond the scope of this book, but Chapters 10 and 11 discuss some of the fundamental experimental design concepts.

Survey responses represent a third type of data source. People being surveyed are asked questions about their beliefs, attitudes, behaviors, and other characteristics. For example, people could be asked which laundry detergent has the best stain-removing abilities. (Such a survey could lead to data that differ from the data collected from the outcomes of the designed experiment of the previous paragraph.) Surveys can be affected by any of the four types of errors that are discussed in Section 1.5.

Observational study results are a fourth data source. A researcher collects data by directly observing a behavior, usually in a natural or neutral setting. Observational studies are a common tool for data collection in business. For example, market researchers use focus groups to elicit unstructured responses to open-ended questions posed by a moderator to a target audience. Observational studies are also commonly used to enhance teamwork or improve the quality of products and services.

Data collected by ongoing business activities are a fifth data source. Such data can be collected from operational and transactional systems that exist in both physical “bricks-and-mortar” and online settings but can also be gathered from secondary sources such as third-party social

LEARN MORE

Read the SHORT TAKES for Chapter 1 for a further discussion about data sources.

media networks and online apps and website services that collect tracking and usage data. For example, a bank might analyze a decade's worth of financial transaction data to identify patterns of fraud, and a marketer might use tracking data to determine the effectiveness of a website.

Sources for “big data” (see Section GS.3) tend to be a mix of primary and secondary sources of this last type. For example, a retailer interested in increasing sales might mine Facebook and Twitter accounts to identify sentiment about certain products or to pinpoint top influencers and then match those data to its own data collected during customer transactions.

Populations and Samples

You collect your data from either a *population* or a *sample*. A **population** consists of all the items or individuals about which you want to reach conclusions. All the GT&M sales transactions for a specific year, all the customers who shopped at GT&M this weekend, all the full-time students enrolled in a college, and all the registered voters in Ohio are examples of populations. In Chapter 3, you will learn that when you analyze data from a population you compute **parameters**.

A **sample** is a portion of a population selected for analysis. The results of analyzing a sample are used to estimate characteristics of the entire population. From the four examples of populations just given, you could select a sample of 200 GT&M sales transactions randomly selected by an auditor for study, a sample of 30 GT&M customers asked to complete a customer satisfaction survey, a sample of 50 full-time students selected for a marketing study, and a sample of 500 registered voters in Ohio contacted via telephone for a political poll. In each of these examples, the transactions or people in the sample represent a portion of the items or individuals that make up the population. In Chapter 3, you will learn that when you analyze data from a sample you compute **statistics**.

Data collection will involve collecting data from a sample when any of the following conditions hold:

- Selecting a sample is less time consuming than selecting every item in the population.
- Selecting a sample is less costly than selecting every item in the population.
- Analyzing a sample is less cumbersome and more practical than analyzing the entire population.

Data Formatting

The data you collect may be formatted in more than one way. For example, suppose that you wanted to collect electronic financial data about a sample of companies. The data you seek to collect could be formatted in any number of ways, including the following:

- Tables of data
- Contents of standard forms
- A continuous data stream, such as a stock ticker
- Messages delivered from social media websites and networks

These examples illustrate that data can exist either in a *structured* or *unstructured* form. **Structured data** is data that follows some organizing principle or plan, typically a repeating pattern. For example, a simple stock ticker is structured because each entry would have the name of a company, the number of shares last traded, the bid price, and percent change in the stock price that the transaction represents. Due to their inherent organization, tables and forms are also structured. In a table, each row contains a set of values for the same columns (i.e., variables), and in a set of forms, each form contains the same set of entries. For example, once we identify that the second column of a table or the second entry on a form contains the family name of an individual, then we know that all entries in the second column of the table or all of the second entries in all copies of the form contain the family name of an individual.

In contrast, **unstructured data** follows no repeating pattern. For example, if five different persons sent you an email message concerning the stock trades of a specific company, that data could be anywhere in the message. You could not reliably count on the name of the company being the first words of each message (as in a stock ticker entry), and the pricing, volume, and

 **Student Tip**

To help remember the difference between a sample and a population, think of a pie. The entire pie represents the population, and the pie slice that you select is the sample.

percent change data could appear in any order. In the Getting Started chapter, *big data* was defined, in part, as data that arrive in a variety of forms, organized and unorganized. You can restate that definition as *big data* exists as both structured and unstructured data.

The ability to handle unstructured data represents an advance in information technology. Chapter 17 discusses business analytics methods that can analyze structured data as well as unstructured data or *semistructured* data. (Think of an application form that contains structured form fills but also contains an unstructured free-response portion.)

With the exception of some of the methods discussed in Chapter 17, the methods taught, and the software techniques used in this book, involve structured data. Your beginning point will always be tabular data, and for many problems and examples, you can begin with that data in the form of a Microsoft Excel or Minitab worksheet that you can download and use (see Appendix C).

Electronic Formats and Encodings Data can exist in more than one electronic format. This affects data formatting as some electronic formats are more immediately usable than others. For example, which data would you like to use: data in an electronic worksheet file or data in a scanned image file that contains one of the worksheet illustrations in this book? Unless you like to do extra “busy work,” you chose the first format because the second format would require you to employ a translation process—perhaps a character-scanning program that can recognize numbers in an image.

Data can also be *encoded* in more than one way, as you may have learned in an information systems course. Different encodings can affect the precision of values for numerical variables, and that can make some data not fully compatible with other data you have collected. While beyond the scope of this book to fully explain, the SHORT TAKES for Chapter 1 includes an experiment that you can perform in either Microsoft Excel or Minitab that illustrates how data encoding can affect precision.

Data Cleaning

Whatever ways you choose to collect data, you may find irregularities in the values you collect such as undefined or impossible values. For a categorical variable, an undefined value would be a value that does not represent one of the categories defined for the variable. For a numerical variable, an impossible value would be a value that falls outside a defined range of possible values for the variable. For a numerical variable without a defined range of possible values, you might also find **outliers**, values that seem excessively different from most of the rest of the values. Such values may or may not be errors, but they demand a second review.

Values that are *missing* are another type of irregularity. A **missing value** is a value that was not able to be collected (and therefore not available to analysis). For example, you would record a nonresponse to a survey question as a missing value. You can represent missing values in Minitab by using an asterisk value for a numerical variable or by using a blank value for a categorical variable, and such values will be properly excluded from analysis. The more limited Excel has no special values that represent a missing value. When using Excel, you must find and then exclude missing values manually.

When you spot an irregularity, you may have to “clean” the data you have collected. Although a full discussion of data cleaning is beyond the scope of this book (see reference 8), you can learn more about the ways you can use Excel or Minitab for data cleaning in the SHORT TAKES for Chapter 1. If you only use the data files designed for use with this book and available online (see Appendix C), you will not need to worry about data cleaning as none of those data files contain any irregularities.

Recoding Variables

After you have collected data, you may discover that you need to reconsider the categories that you have defined for a categorical variable or that you need to transform a numerical variable into a categorical variable by assigning the individual numeric data values to one of several groups. In either case, you can define a **recoded variable** that supplements or replaces the original variable in your analysis.

For example, having defined the variable student class designation to be one of the four categories shown in Table 1.3 on page 16, you realize that you are more interested in investigating the differences between lowerclassmen (defined as freshman or sophomore) and upperclassmen (junior or senior). You can create a new variable UpperLower and assign the value Upper if a student is a junior or senior and assign the value Lower if the student is a freshman or sophomore.

When recoding variables, be sure that the category definitions cause each data value to be placed in one and only one category, a property known as being **mutually exclusive**. Also ensure that the set of categories you create for the new, recoded variables include all the data values being recoded, a property known as being **collectively exhaustive**. If you are recoding a categorical variable, you can preserve one or more of the original categories, as long as your recodings are both mutually exclusive and collectively exhaustive.

When recoding numerical variables, pay particular attention to the operational definitions of the categories you create for the recoded variable, especially if the categories are not self-defining ranges. For example, while the recoded categories Under 12, 12–20, 21–34, 35–54, and 55 and Over are self-defining for age, the categories Child, Youth, Young Adult, Middle Aged, and Senior need their own operational definitions.

Problems for Section 1.3

APPLYING THE CONCEPTS

1.12 The Data and Story Library (DASL) is an online library of data files and stories that illustrate the use of basic statistical methods. Visit lib.stat.cmu.edu/index.php, click DASL, and explore a data set of interest to you. Which of the five sources of data best describes the sources of the data set you selected?

1.13 Visit the website of the Gallup organization at www.gallup.com. Read today's top story. What type of data source is the top story based on?

1.14 Visit the website of the Pew Research organization at www.pewresearch.org. Read today's top story. What type of data source is the top story based on?

1.15 Transportation engineers and planners want to address the dynamic properties of travel behavior by describing in detail the driving characteristics of drivers over the course of a month. What type of data collection source do you think the transportation engineers and planners should use?

1.16 Visit the opening page of the Statistics Portal "Statista" at (statista.com). Examine the "CHART OF THE DAY" panel on the page. What type of data source is the information presented here based on?

1.4 Types of Sampling Methods

When you collect data by selecting a sample, you begin by defining the **frame**. The frame is a complete or partial listing of the items that make up the population from which the sample will be selected. Inaccurate or biased results can occur if a frame excludes certain groups, or portions of the population. Using different frames to collect data can lead to different, even opposite, conclusions.

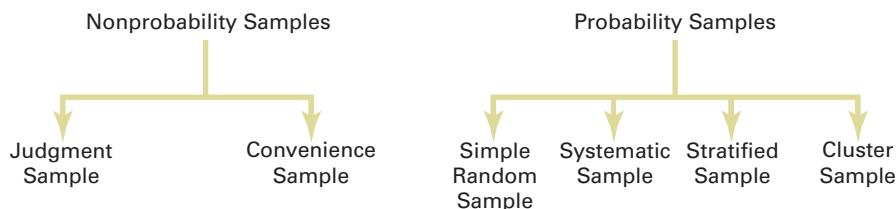
Using your frame, you select either a nonprobability sample or a probability sample. In a **nonprobability sample**, you select the items or individuals without knowing their probabilities of selection. In a **probability sample**, you select items based on known probabilities. Whenever possible, you should use a probability sample as such a sample will allow you to make inferences about the population being analyzed.

Nonprobability samples can have certain advantages, such as convenience, speed, and low cost. Such samples are typically used to obtain informal approximations or as small-scale initial or pilot analyses. However, because the theory of statistical inference depends on probability sampling, nonprobability samples *cannot be used* for statistical inference and this more than offsets those advantages in more formal analyses.

Figure 1.1 shows the subcategories of the two types of sampling. A nonprobability sample can be either a convenience sample or a judgment sample. To collect a **convenience sample**, you select items that are easy, inexpensive, or convenient to sample. For example, in a warehouse of stacked items, selecting only the items located on the tops of each stack and within

FIGURE 1.1

Types of samples



easy reach would create a convenience sample. So, too, would be the responses to surveys that the websites of many companies offer visitors. While such surveys can provide large amounts of data quickly and inexpensively, the convenience samples selected from these responses will consist of self-selected website visitors. (Read the Think About This essay on page 26 for a related story.)

To collect a **judgment sample**, you collect the opinions of preselected experts in the subject matter. Although the experts may be well informed, you cannot generalize their results to the population.

The types of probability samples most commonly used include simple random, systematic, stratified, and cluster samples. These four types of probability samples vary in terms of cost, accuracy, and complexity, and they are the subject of the rest of this section.

Simple Random Sample

In a **simple random sample**, every item from a frame has the same chance of selection as every other item, and every sample of a fixed size has the same chance of selection as every other sample of that size. Simple random sampling is the most elementary random sampling technique. It forms the basis for the other random sampling techniques. However, simple random sampling has its disadvantages. Its results are often subject to more variation than other sampling methods. In addition, when the frame used is very large, carrying out a simple random sample may be time consuming and expensive.

With simple random sampling, you use n to represent the sample size and N to represent the frame size. You number every item in the frame from 1 to N . The chance that you will select any particular member of the frame on the first selection is $1/N$.

You select samples with replacement or without replacement. **Sampling with replacement** means that after you select an item, you return it to the frame, where it has the same probability of being selected again. Imagine that you have a fishbowl containing N business cards, one card for each person. On the first selection, you select the card for Grace Kim. You record pertinent information and replace the business card in the bowl. You then mix up the cards in the bowl and select a second card. On the second selection, Grace Kim has the same probability of being selected again, $1/N$. You repeat this process until you have selected the desired sample size, n .

Typically, you do not want the same item or individual to be selected again in a sample. **Sampling without replacement** means that once you select an item, you cannot select it again. The chance that you will select any particular item in the frame—for example, the business card for Grace Kim—on the first selection is $1/N$. The chance that you will select any card not previously chosen on the second selection is now 1 out of $N - 1$. This process continues until you have selected the desired sample of size n .

When creating a simple random sample, you should avoid the “fishbowl” method of selecting a sample because this method lacks the ability to thoroughly mix the cards and, therefore, randomly select a sample. You should use a more rigorous selection method.

One such method is to use a **table of random numbers**, such as Table E.1 in Appendix E, for selecting the sample. A table of random numbers consists of a series of digits listed in a randomly generated sequence. To use a random number table for selecting a sample, you first need to assign code numbers to the individual items of the frame. Then you generate the random sample by reading the table of random numbers and selecting those individuals from the frame whose assigned code numbers match the digits found in the table. Because the number system uses 10 digits ($0, 1, 2, \dots, 9$), the chance that you will randomly generate any

LEARN MORE

Learn to use a table of random numbers to select a simple random sample in a Chapter 1 online section.

particular digit is equal to the probability of generating any other digit. This probability is 1 out of 10. Hence, if you generate a sequence of 800 digits, you would expect about 80 to be the digit 0, 80 to be the digit 1, and so on. Because every digit or sequence of digits in the table is random, the table can be read either horizontally or vertically. The margins of the table designate row numbers and column numbers. The digits themselves are grouped into sequences of five in order to make reading the table easier.

Systematic Sample

In a **systematic sample**, you partition the N items in the frame into n groups of k items, where

$$k = \frac{N}{n}$$

You round k to the nearest integer. To select a systematic sample, you choose the first item to be selected at random from the first k items in the frame. Then, you select the remaining $n - 1$ items by taking every k th item thereafter from the entire frame.

If the frame consists of a list of prenumbered checks, sales receipts, or invoices, taking a systematic sample is faster and easier than taking a simple random sample. A systematic sample is also a convenient mechanism for collecting data from membership directories, electoral registers, class rosters, and consecutive items coming off an assembly line.

To take a systematic sample of $n = 40$ from the population of $N = 800$ full-time employees, you partition the frame of 800 into 40 groups, each of which contains 20 employees. You then select a random number from the first 20 individuals and include every twentieth individual after the first selection in the sample. For example, if the first random number you select is 008, your subsequent selections are 028, 048, 068, 088, 108, . . . , 768, and 788.

Simple random sampling and systematic sampling are simpler than other, more sophisticated, probability sampling methods, but they generally require a larger sample size. In addition, systematic sampling is prone to selection bias that can occur when there is a pattern in the frame. To overcome the inefficiency of simple random sampling and the potential selection bias involved with systematic sampling, you can use either stratified sampling methods or cluster sampling methods.

Stratified Sample

LEARN MORE

Learn how to select a stratified sample in a Chapter 1 online section.

In a **stratified sample**, you first subdivide the N items in the frame into separate subpopulations, or **strata**. A stratum is defined by some common characteristic, such as gender or year in school. You select a simple random sample within each of the strata and combine the results from the separate simple random samples. Stratified sampling is more efficient than either simple random sampling or systematic sampling because you are ensured of the representation of items across the entire population. The homogeneity of items within each stratum provides greater precision in the estimates of underlying population parameters. In addition, stratified sampling enables you to reach conclusions about each strata in the frame. However, using a stratified sample requires that you can determine the variable(s) on which to base the stratification and can also be expensive to implement.

Cluster Sample

In a **cluster sample**, you divide the N items in the frame into clusters that contain several items. **Clusters** are often naturally occurring groups, such as counties, election districts, city blocks, households, or sales territories. You then take a random sample of one or more clusters and study all items in each selected cluster.

Cluster sampling is often more cost-effective than simple random sampling, particularly if the population is spread over a wide geographic region. However, cluster sampling often requires a larger sample size to produce results as precise as those from simple random sampling or stratified sampling. A detailed discussion of systematic sampling, stratified sampling, and cluster sampling procedures can be found in references 2, 4, and 5.

Problems for Section 1.4

LEARNING THE BASICS

1.17 For a population containing $N = 902$ individuals, what code number would you assign for

- the first person on the list?
- the fortieth person on the list?
- the last person on the list?

1.18 For a population of $N = 902$, verify that by starting in row 05, column 01 of the table of random numbers (Table E.1), you need only six rows to select a sample of $N = 60$ without replacement.

1.19 Given a population of $N = 93$, starting in row 29, column 01 of the table of random numbers (Table E.1), and reading across the row, select a sample of $N = 15$

- without replacement.
- with replacement.

APPLYING THE CONCEPTS

1.20 For a study that consists of personal interviews with participants (rather than mail or phone surveys), explain why simple random sampling might be less practical than some other sampling methods.

1.21 You want to select a random sample of $n = 1$ from a population of three items (which are called *A*, *B*, and *C*). The rule for selecting the sample is as follows: Flip a coin; if it is heads, pick item *A*; if it is tails, flip the coin again; this time, if it is heads, choose *B*; if it is tails, choose *C*. Explain why this is a probability sample but not a simple random sample.

1.22 A population has four members (called *A*, *B*, *C*, and *D*). You would like to select a random sample of $n = 2$, which you decide to do in the following way: Flip a coin; if it is heads, the sample will be items *A* and *B*; if it is tails, the sample will be items *C* and *D*. Although this is a random sample, it is not a simple random sample. Explain why. (Compare the procedure described in Problem 1.21 with the procedure described in this problem.)

1.23 The registrar of a college with a population of $N = 4,000$ full-time students is asked by the president to conduct a survey to measure satisfaction with the quality of life on campus. The following table contains a breakdown of the 4,000 registered full-time students, by gender and class designation:

Class Designation

Gender	Fr.	So.	Jr.	Sr.	Total
Female	700	520	500	480	2,200
Male	560	460	400	380	1,800
Total	1,260	980	900	860	4,000

The registrar intends to take a probability sample of $n = 200$ students and project the results from the sample to the entire population of full-time students.

- If the frame available from the registrar's files is an alphabetical listing of the names of all $N = 4,000$ registered full-time students, what type of sample could you take? Discuss.
- What is the advantage of selecting a simple random sample in (a)?
- What is the advantage of selecting a systematic sample in (a)?
- If the frame available from the registrar's files is a list of the names of all $N = 4,000$ registered full-time students compiled from eight separate alphabetical lists, based on the gender and class designation breakdowns shown in the class designation table, what type of sample should you take? Discuss.
- Suppose that each of the $N = 4,000$ registered full-time students lived in one of the 10 campus dormitories. Each dormitory accommodates 400 students. It is college policy to fully integrate students by gender and class designation in each dormitory. If the registrar is able to compile a listing of all students by dormitory, explain how you could take a cluster sample.

✓ SELF TEST **1.24** Prenumbered sales invoices are kept in a sales journal. The invoices are numbered from 0001 to 5000.

- Beginning in row 16, column 01, and proceeding horizontally in a table of random numbers (Table E.1), select a simple random sample of 50 invoice numbers.
- Select a systematic sample of 50 invoice numbers. Use the random numbers in row 20, columns 05–07, as the starting point for your selection.
- Are the invoices selected in (a) the same as those selected in (b)? Why or why not?

1.25 Suppose that 5,000 sales invoices are separated into four strata. Stratum 1 contains 50 invoices, stratum 2 contains 500 invoices, stratum 3 contains 1,000 invoices, and stratum 4 contains 3,450 invoices. A sample of 500 sales invoices is needed.

- What type of sampling should you do? Why?
- Explain how you would carry out the sampling according to the method stated in (a).
- Why is the sampling in (a) not simple random sampling?

1.5 Types of Survey Errors

As you learned in Section 1.3, responses from a survey represent a source of data. Nearly every day, you read or hear about survey or opinion poll results in newspapers, on the Internet, or on radio or television. To identify surveys that lack objectivity or credibility, you must critically evaluate what you read and hear by examining the validity of the survey results.

First, you must evaluate the purpose of the survey, why it was conducted, and for whom it was conducted.

The second step in evaluating the validity of a survey is to determine whether it was based on a probability or nonprobability sample (as discussed in Section 1.4). You need to remember that the only way to make valid statistical inferences from a sample to a population is by using a probability sample. Surveys that use nonprobability sampling methods are subject to serious biases that may make the results meaningless.

Even when surveys use probability sampling methods, they are subject to four types of potential survey errors:

- Coverage error
- Nonresponse error
- Sampling error
- Measurement error

Well-designed surveys reduce or minimize these four types of errors, often at considerable cost.

Coverage Error

The key to proper sample selection is having an adequate frame. **Coverage error** occurs if certain groups of items are excluded from the frame so that they have no chance of being selected in the sample or if items are included from outside the frame. Coverage error results in a **selection bias**. If the frame is inadequate because certain groups of items in the population were not properly included, any probability sample selected will provide only an estimate of the characteristics of the frame, not the *actual* population.

Nonresponse Error

Not everyone is willing to respond to a survey. **Nonresponse error** arises from failure to collect data on all items in the sample and results in a **nonresponse bias**. Because you cannot always assume that persons who do not respond to surveys are similar to those who do, you need to follow up on the nonresponses after a specified period of time. You should make several attempts to convince such individuals to complete the survey and possibly offer an incentive to participate. The follow-up responses are then compared to the initial responses in order to make valid inferences from the survey (see references 2, 4, and 5). The mode of response you use, such as face-to-face interview, telephone interview, paper questionnaire, or computerized questionnaire, affects the rate of response. Personal interviews and telephone interviews usually produce a higher response rate than do mail surveys—but at a higher cost.

Sampling Error

When conducting a probability sample, chance dictates which individuals or items will or will not be included in the sample. **Sampling error** reflects the variation, or “chance differences,” from sample to sample, based on the probability of particular individuals or items being selected in the particular samples.

When you read about the results of surveys or polls in newspapers or on the Internet, there is often a statement regarding a margin of error, such as “the results of this poll are expected to be within ± 4 percentage points of the actual value.” This **margin of error** is the sampling error. You can reduce sampling error by using larger sample sizes. Of course, doing so increases the cost of conducting the survey.

Measurement Error

In the practice of good survey research, you design surveys with the intention of gathering meaningful and accurate information. Unfortunately, the survey results you get are often only a proxy for the ones you really desire. Unlike height or weight, certain information about behaviors and psychological states is impossible or impractical to obtain directly.

When surveys rely on self-reported information, the mode of data collection, the respondent to the survey, and/or the survey itself can be possible sources of **measurement error**.

Satisficing, social desirability, reading ability, and/or interviewer effects can be dependent on the mode. The social desirability bias or cognitive/memory limitations of a respondent can affect the results. And vague questions, double-barreled questions that ask about multiple issues but require a single response, or questions that ask the respondent to report something that occurs over time but fail to clearly define the extent of time about which the question asks (the reference period) are some of the survey flaws that can cause errors.

To minimize measurement error, you need to standardize survey administration and respondent understanding of questions, but there are many barriers to this (see references 1, 3, and 10).

Ethical Issues About Surveys

Ethical considerations arise with respect to the four types of survey error. Coverage error can result in selection bias and becomes an ethical issue if particular groups or individuals are purposely excluded from the frame so that the survey results are more favorable to the survey's sponsor. Nonresponse error can lead to nonresponse bias and becomes an ethical issue if the sponsor knowingly designs the survey so that particular groups or individuals are less likely than others to respond. Sampling error becomes an ethical issue if the findings are purposely presented without reference to sample size and margin of error so that the sponsor can promote a viewpoint that might otherwise be inappropriate. Measurement error can become an ethical issue in one of three ways: (1) a survey sponsor chooses leading questions that guide the respondent in a particular direction; (2) an interviewer, through mannerisms and tone, purposely makes a respondent obligated to please the interviewer or otherwise guides the respondent in a particular direction; or (3) a respondent willfully provides false information.

Ethical issues also arise when the results of nonprobability samples are used to form conclusions about the entire population. When you use a nonprobability sampling method, you need to explain the sampling procedures and state that the results cannot be generalized beyond the sample.

THINK ABOUT THIS

New Media Surveys/Old Sampling Problems

Imagine that you work for a software distributor that has decided to create a "customer experience improvement program" to record how your customers are using your products, with the goal of using the collected data to make product enhancements. Or say that you are an editor of an online news website who decides to create an instant poll to ask website visitors about important political issues. Or you're a marketer of products aimed at a specific demographic and decide to use a social networking site to collect consumer feedback. What might you have in common with a *dead-tree* publication that went out of business over 70 years ago?

By 1932, before there was ever an Internet—or even commercial television—a "straw poll" conducted by the magazine *Literary Digest* had successfully predicted five U.S. presidential elections in a row. For the 1936 election, the magazine promised its largest poll ever and sent about 10 million ballots to people all across the country. After receiving and tabulating more than

2.3 million ballots, the *Digest* confidently proclaimed that Alf Landon would be an easy winner over Franklin D. Roosevelt. As things turned out, FDR won in a landslide, with Landon receiving the fewest electoral votes in U.S. history. The reputation of *Literary Digest* was ruined; the magazine would cease publication less than two years later.

The failure of the *Literary Digest* poll was a watershed event in the history of sample surveys and polls. This failure refuted the notion that the larger the sample is, the better. (Remember this the next time someone complains about a political survey's "small" sample size.) The failure opened the door to new and more modern methods of sampling discussed in this chapter. Today's Gallup polls of political opinion (www.gallup.com) or GfK Roper Reports about consumer behavior (www.gfkamerica.com/practice_areas/roper_consulting) arose, in part, due to this failure. George Gallup, the "Gallup" of the poll, and Elmo Roper, of the eponymous reports, both first gained

widespread public notice for their correct "scientific" predictions of the 1936 election.

The failed *Literary Digest* poll became fodder for several postmortems, and the reason for the failure became almost an urban legend. Typically, the explanation is coverage error: The ballots were sent mostly to "rich people," and this created a frame that excluded poorer citizens (presumably more inclined to vote for the Democrat Roosevelt than the Republican Landon). However, later analyses suggest that this was not true; instead, low rates of response (2.3 million ballots represented less than 25% of the ballots distributed) and/or nonresponse error (Roosevelt voters were less likely to mail in a ballot than Landon voters) were significant reasons for the failure (see reference 9).

When Microsoft revealed its Office Ribbon for Office 2007, a program manager explained how Microsoft had applied data collected from its "Customer Experience Improvement Program" to the user

interface redesign. This led others to speculate that the data were biased toward beginners—who might be less likely to decline participation in the program—and that, in turn, had led Microsoft to create a user interface that ended up perplexing more experienced users. This was another case of nonresponse error!

The editor's instant poll mentioned earlier is targeted to the visitors of the online news website, and the social network-based survey is aimed at "friends" of a product; such polls can also suffer

from nonresponse error, and this fact is often overlooked by users of these new media. Often, marketers extol how much they "know" about survey respondents, thanks to data that can be collected from a social network community. But no amount of information about the respondents can tell marketers who the nonrespondents are. Therefore, new media surveys fall prey to the same old type of error that may have been fatal to *Literary Digest* way back when.

Today, companies establish formal surveys based on probability sampling and go to great lengths—and spend large sums—to deal with coverage error, nonresponse error, sampling error, and measurement error. Instant polling and tell-a-friend surveys can be interesting and fun, but they are not replacements for the methods discussed in this chapter.

Problems for Section 1.5

APPLYING THE CONCEPTS

1.26 A survey indicates that the vast majority of college students own their own personal computers. What information would you want to know before you accepted the results of this survey?

1.27 A simple random sample of $n = 300$ full-time employees is selected from a company list containing the names of all $N = 5,000$ full-time employees in order to evaluate job satisfaction.

- a. Give an example of possible coverage error.
- b. Give an example of possible nonresponse error.
- c. Give an example of possible sampling error.
- d. Give an example of possible measurement error.

 **1.28** The results of a 2013 Adobe Systems study on retail apps and buying habits reveal insights on perceptions and attitudes toward mobile shopping using retail apps and browsers, providing new direction for retailers to develop their digital publishing strategies (adobe.ly/11gt8Rq). Increased consumer interest in using shopping applications means retailers must adapt to meet the rising expectations for specialized mobile shopping experiences. The results indicate that tablet users (55%) are almost twice as likely as smartphone users (28%) to use their device to purchase products and services. The findings also reveal that retail and catalog apps are rapidly catching up to mobile

browsers as a viable shopping channel: nearly half of all mobile shoppers are interested in using apps instead of a mobile browser (45% of tablet shoppers and 49% of smartphone shoppers). The research is based on an online survey with a sample of 1,003 consumers. Identify potential concerns with coverage, nonresponse, sampling, and measurement errors.

1.29 A recent PwC Supply Global Chain survey indicated that companies that acknowledge the supply chain as a strategic asset achieve 70% higher performance (pwc.to/VaFpGz). The "Leaders" in the survey point to next-generation supply chains, which are fast, flexible, and responsive. They are more concerned with skills that separate a company from the crowd: 51% say differentiating capabilities is the real key to success. What additional information would you want to know about the survey before you accepted the results of the study?

1.30 A recent survey points to a next generation of consumers seeking a more mobile TV experience. The 2013 KPMG International Consumer Media Behavior study found that while TV is still the most popular media activity with 88% of U.S. consumers watching TV, a relatively high proportion of U.S. consumers, 14%, now prefer to watch TV via their mobile device or tablet for greater flexibility (bit.ly/Wb8Jv9). What additional information would you want to know about the survey before you accepted the results of the study?

USING STATISTICS

Beginning of the End... Revisited

The analysts charged by GT&M CEO Emma Levia to identify, define, and collect the data that would be helpful in setting a price for Whitney Wireless have completed their task. The group has identified a number of variables to analyze. In the course of doing this work, the group realized that most of the variables to study would be discrete numerical variables based on data that (ac)counts the financials of the business. These data would mostly be from the

primary source of the business itself, but some supplemental variables

about economic conditions and other factors that might affect the long-term prospects of the business might come from a secondary data source, such as an economic agency.



Tyler Olson/Shutterstock

The group foresaw that examining several categorical variables related to the customers of both GT&M and Whitney Wireless would be necessary. The group discovered that the affinity (“shopper’s card”) programs of both firms had already collected demographic data of interest when customers enrolled in those programs. That primary source, when combined

with secondary data gleaned from the social media networks to which the business belongs, might prove useful in getting a rough approximation of the profile of a typical customer that might be interested in doing business with an “A-to-Z” electronics retailer.

SUMMARY

In this chapter, you learned about the various types of variables used in business and their measurement scales. In addition, you learned about different methods of collecting data, several statistical sampling methods, and issues

involved in taking samples. In the next two chapters, you will study a variety of tables and charts and descriptive measures that are used to present and analyze data.

REFERENCES

1. Biemer, P. B., R. M. Graves, L. E. Lyberg, A. Mathiowetz, and S. Sudman. *Measurement Errors in Surveys*. New York: Wiley Interscience, 2004.
2. Cochran, W. G. *Sampling Techniques*, 3rd ed. New York: Wiley, 1977.
3. Fowler, F. J. *Improving Survey Questions: Design and Evaluation, Applied Special Research Methods Series*, Vol. 38. Thousand Oaks, CA: Sage Publications, 1995.
4. Groves R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. *Survey Methodology*, 2nd ed. New York: John Wiley, 2009.
5. Lohr, S. L. *Sampling Design and Analysis*, 2nd ed. Boston, MA: Brooks/Cole Cengage Learning, 2010.
6. Microsoft Excel 2013. Redmond, WA: Microsoft Corporation, 2012.
7. Minitab Release 16. State College, PA: Minitab, Inc., 2010.
8. Osbourne, J. *Best Practices in Data Cleaning*. Thousand Oaks, CA: Sage Publications, 2012.
9. Squire, P. “Why the 1936 Literary Digest Poll Failed.” *Public Opinion Quarterly* 52 (1988): 125–133.
10. Sudman, S., N. M. Bradburn, and N. Schwarz. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco, CA: Jossey-Bass, 1993.

KEY TERMS

categorical variable 14
 cluster 23
 cluster sample 23
 collect 14
 collectively exhaustive 21
 continuous variable 15
 convenience sample 21
 coverage error 25
 define 14
 discrete variable 15
 frame 21
 interval scale 16
 judgment sample 22
 margin of error 25
 measurement error 25
 measurement scale 15

missing value 20
 mutually exclusive 21
 nominal scale 15
 nonprobability sample 21
 nonresponse bias 25
 nonresponse error 25
 numerical variable 15
 operational definition 14
 ordinal scale 15
 outlier 20
 parameter 19
 population 19
 primary data source 18
 probability sample 21
 qualitative variable 14
 quantitative variable 15

ratio scale 17
 recoded variable 20
 sample 19
 sampling error 25
 sampling with replacement 22
 sampling without replacement 22
 secondary data source 18
 selection bias 25
 simple random sample 22
 statistics 19
 strata 23
 stratified sample 23
 structured data 19
 systematic sample 23
 table of random numbers 22
 unstructured data 19

CHECKING YOUR UNDERSTANDING

- 1.31** What is the difference between a sample and a population?
- 1.32** What is the difference between a statistic and a parameter?
- 1.33** What is the difference between a categorical variable and a numerical variable?
- 1.34** What is the difference between a discrete numerical variable and a continuous numerical variable?

- 1.35** What is the difference between a nominal scaled variable and an ordinal scaled variable?
- 1.36** What is the difference between an interval-scaled variable and a ratio-scaled variable?
- 1.37** What is the difference between probability sampling and non-probability sampling?

CHAPTER REVIEW PROBLEMS

1.38 Visit the official website for either Excel (www.office.microsoft.com/excel) or Minitab (www.minitab.com/products/minitab). Read about the program you chose and then think about the ways the program could be useful in statistical analysis.

1.39 Results of a 2013 Adobe Systems study on retail apps and buying habits reveals insights on perceptions and attitudes toward mobile shopping using retail apps and browsers, providing new direction for retailers to develop their digital publishing strategies. Increased consumer interest in using shopping applications means retailers must adapt to meet the rising expectations for specialized mobile shopping experiences. The results indicate that tablet users (55%) are almost twice as likely as smartphone users (28%) to use their device to purchase products and services. The findings also reveal that retail and catalog apps are rapidly catching up to mobile browsers as a viable shopping channel: Nearly half of all mobile shoppers are interested in using apps instead of a mobile browser (45% of tablet shoppers and 49% of smartphone shoppers). The research is based on an online survey with a sample of 1,003 18–54 year olds who currently own a smartphone and/or tablet; it includes consumers who use and do not use these devices to shop (adobe.ly/11gt8Rq).

- Describe the population of interest.
- Describe the sample that was collected.
- Describe a parameter of interest.
- Describe the statistic used to estimate the parameter in (c).

1.40 The Gallup organization releases the results of recent polls at its website, www.gallup.com. Visit this site and read an article of interest.

- Describe the population of interest.
- Describe the sample that was collected.
- Describe a parameter of interest.
- Describe the statistic used to estimate the parameter in (c).

1.41 A recent PwC Supply Global Chain survey indicated that companies that acknowledge the supply chain as a strategic asset achieve 70% higher performance. The “Leaders” in the survey point to next-generation supply chains, which are fast, flexible, and responsive. They are more concerned with skills that separate a company from the crowd: 51% say differentiating capabilities is the real key to success (pwc.to/VaFpGz). The results are based on a survey of 503 supply chain executives in a wide range of industries representing a mix of company sizes from across three global regions: Asia, Europe, and the Americas.

- Describe the population of interest.
- Describe the sample that was collected.
- Describe a parameter of interest.
- Describe the statistic used to estimate the parameter in (c).

1.42 The Data and Story Library (DASL) is an online library of data files and stories that illustrate the use of basic statistical methods. Visit lib.stat.cmu.edu/index.php, click DASL, and explore a data set of interest to you.

- Describe a variable in the data set you selected.
- Is the variable categorical or numerical?
- If the variable is numerical, is it discrete or continuous?

1.43 Download and examine the U.S. Census Bureau’s “Business and Professional Classification Survey (SQ-CLASS),” available through the **Get Help with Your Form** link at www.census.gov/econ/.

- Give an example of a categorical variable included in the survey.
- Give an example of a numerical variable included in the survey.

1.44 Three professors examined awareness of four widely disseminated retirement rules among employees at the University of Utah. These rules provide simple answers to questions about retirement planning (R. N. Mayer, C. D. Zick, and M. Glaitle, “Public Awareness of Retirement Planning Rules of Thumb,” *Journal of Personal Finance*, 2011 10(1), 12–35). At the time of the investigation, there were approximately 10,000 benefited employees, and 3,095 participated in the study. Demographic data collected on these 3,095 employees included gender, age (years), education level (years completed), marital status, household income (\$), and employment category.

- Describe the population of interest.
- Describe the sample that was collected.
- Indicate whether each of the demographic variables mentioned is categorical or numerical.

1.45 A manufacturer of cat food is planning to survey households in the United States to determine purchasing habits of cat owners. Among the variables to be collected are the following:

- The primary place of purchase for cat food
 - Whether dry or moist cat food is purchased
 - The number of cats living in the household
 - Whether any cat living in the household is pedigreed
- For each of the four items listed, indicate whether the variable is categorical or numerical. If it is numerical, is it discrete or continuous?
 - Develop five categorical questions for the survey.
 - Develop five numerical questions for the survey.

CASES FOR CHAPTER 1

Managing Ashland MultiComm Services

Ashland MultiComm Services (AMS) provides high-quality communications networks in the Greater Ashland area. AMS traces its roots to Ashland Community Access Television (ACATV), a small company that redistributed the broadcast television signals from nearby major metropolitan areas but has evolved into a provider of a wide range of broadband services for residential customers.

AMS offers subscription-based services for digital cable video programming, local and long-distance telephone services, and high-speed Internet access. Recently, AMS has faced competition from other network providers that have expanded into the Ashland area. AMS has also seen decreases in the number of new digital cable installations and the rate of digital cable renewals.

AMS management believes that a combination of increased promotional expenditures, adjustment in subscription fees, and improved customer service will allow AMS to successfully face the competition from other network providers. However, AMS management worries about the possible effects that new Internet-based methods of program delivery may have had on their digital cable business. They decide that they need to conduct some research and organize

a team of research specialists to examine the current status of the business and the marketplace in which it competes.

The managers suggest that the research team examine the company's own historical data for number of subscribers, revenues, and subscription renewal rates for the past few years. They direct the team to examine year-to-date data as well, as the managers suspect that some of the changes they have seen have been a relatively recent phenomena.

1. What type of data source would the company's own historical data be? Identify other possible data sources that the research team might use to examine the current marketplace for residential broadband services in a city such as Ashland.
2. What type of data collection techniques might the team employ?
3. In their suggestions and directions, the AMS managers have named a number of possible variables to study, but offered no operational definitions for those variables. What types of possible misunderstandings could arise if the team and managers do not first properly define each variable cited?

CardioGood Fitness

CardioGood Fitness is a developer of high-quality cardiovascular exercise equipment. Its products include treadmills, fitness bikes, elliptical machines, and e-glides. CardioGood Fitness looks to increase the sales of its treadmill products and has hired The AdRight Agency, a small advertising firm, to create and implement an advertising program. The AdRight Agency plans to identify particular market segments that are most likely to buy their clients' goods and services and then locates advertising outlets that will reach that market group. This activity includes collecting data on clients' actual sales and on the customers who make the purchases, with the goal of determining whether there is a distinct profile of the typical customer for a particular product or service. If a distinct profile emerges, efforts are made to match that profile to advertising outlets known to reflect the particular profile, thus targeting advertising directly to high-potential customers.

CardioGood Fitness sells three different lines of treadmills. The TM195 is an entry-level treadmill. It is as dependable as other models offered by CardioGood Fitness, but with fewer programs and features. It is suitable for individuals who thrive on minimal programming and the desire

for simplicity to initiate their walk or hike. The TM195 sells for \$1,500.

The middle-line TM498 adds to the features of the entry-level model two user programs and up to 15% elevation upgrade. The TM498 is suitable for individuals who are walkers at a transitional stage from walking to running or midlevel runners. The TM498 sells for \$1,750.

The top-of-the-line TM798 is structurally larger and heavier and has more features than the other models. Its unique features include a bright blue backlit LCD console, quick speed and incline keys, a wireless heart rate monitor with a telemetric chest strap, remote speed and incline controls, and an anatomical figure that specifies which muscles are minimally and maximally activated. This model features a nonfolding platform base that is designed to handle rigorous, frequent running; the TM798 is therefore appealing to someone who is a power walker or a runner. The selling price is \$2,500.

As a first step, the market research team at AdRight is assigned the task of identifying the profile of the typical customer for each treadmill product offered by CardioGood Fitness. The market research team decides to investigate

whether there are differences across the product lines with respect to customer characteristics. The team decides to collect data on individuals who purchased a treadmill at a CardioGood Fitness retail store during the prior three months.

The team decides to use both business transactional data and the results of a personal profile survey that every purchaser completes as their sources of data. The team identifies the following customer variables to study: product purchased—TM195, TM498, or TM798; gender; age, in years; education, in years; relationship status, single or partnered; annual household income (\$); average number

of times the customer plans to use the treadmill each week; average number of miles the customer expects to walk/run each week; and self-rated fitness on an 1-to-5 ordinal scale, where 1 is poor shape and 5 is excellent shape. For this set of variables:

1. Which variables in the survey are categorical?
2. Which variables in the survey are numerical?
3. Which variables are discrete numerical variables?

Clear Mountain State Student Surveys

1. The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students who attend CMSU. They create and distribute a survey of 14 questions and receive responses from 62 undergraduates (stored in [UndergradSurvey](#)). Download (see Appendix C) and review the survey document **CMUndergradSurvey.pdf**. For each question asked in the survey, determine whether the variable is categorical or numerical. If you determine that the variable is numerical, identify whether it is discrete or continuous.
2. The dean of students at CMSU has learned about the undergraduate survey and has decided to undertake a similar survey for graduate students at CMSU. She creates and distributes a survey of 14 questions and receives responses from 44 graduate students (stored in [GradSurvey](#)). Download (see Appendix C) and review the survey document **CMSGradSurvey.pdf**. For each question asked in the survey, determine whether the variable is categorical or numerical. If you determine that the variable is numerical, identify whether it is discrete or continuous.

Learning with the Digital Cases

As you have already learned in this book, decision makers use statistical methods to help analyze data and communicate results. Every day, somewhere, someone misuses these techniques either by accident or intentional choice. Identifying and preventing such misuses of statistics is an important responsibility for all managers. The Digital Cases give you the practice you need to help develop the skills necessary for this important task.

Each chapter's Digital Case tests your understanding of how to apply an important statistical concept taught in the chapter. For each case, you review the contents of one or more electronic documents, which may contain internal and confidential information to an organization as well as publicly stated facts and claims, seeking to identify and correct misuses of statistics. Unlike in a traditional case study, but like in many business situations, not all of the information you encounter will be relevant to your task, and you may occasionally discover conflicting information that you have to resolve in order to complete the case.

To assist your learning, each Digital Case begins with a learning objective and a summary of the problem or issue at hand. Each case directs you to the information necessary to reach your own conclusions and to answer the case

questions. Many cases, such as the sample case worked out next, extend a chapter's Using Statistics scenario. You can download digital case files for later use or retrieve them online from a MyStatLab course for this book, as explained in Appendix C.

SAMPLE DIGITAL CASE

To illustrate learning with a Digital Case, open the Digital Case file **WhitneyWireless.pdf** that contains summary information about the Whitney Wireless business. Recall from the Using Statistics scenario for this chapter that Good Tunes & More (GT&M) is a retailer seeking to expand by purchasing Whitney Wireless, a small chain that sells mobile media devices. Apparently, from the claim on the title page, this business is celebrating its “best sales year ever.”

Review the **Who We Are, What We Do, and What We Plan to Do** sections on the second page. Do these sections contain any useful information? What *questions* does this passage raise? Did you notice that while many facts are presented, no data that would support the claim of “best sales year ever” are presented? And were those mobile “mobile-mobiles” used solely for promotion? Or did they generate

any sales? Do you think that a talk-with-your-mouth-full event, however novel, would be a success?

Continue to the third page and the **Our Best Sales Year Ever!** section. How would you support such a claim? With a table of numbers? Remarks attributed to a knowledgeable source? Whitney Wireless has used a chart to present “two years ago” and “latest twelve months” sales data by category. Are there any problems with what the company has done? *Absolutely!*

First, note that there are no scales for the symbols used, so you cannot know what the actual sales volumes are. In fact, as you will learn in Section 2.7, charts that incorporate icons as shown on the third page are considered examples of *chartjunk* and would never be used by people seeking to properly visualize data. The use of chartjunk symbols creates the impression that unit sales data are being presented. If the data are unit sales, does such data best support the claim being made, or would something else, such as dollar volumes, be a better indicator of sales at the retailer?

For the moment, let’s assume that unit sales are being visualized. What are you to make of the second row, in which the three icons on the right side are much wider than the three on the left? Does that row represent a newer (wider) model being sold or a greater sales volume? Examine the fourth row. Does that row represent a decline in sales

or an increase? (Do two partial icons represent more than one whole icon?) As for the fifth row, what are we to think? Is a black icon worth more than a red icon or vice versa?

At least the third row seems to tell some sort of tale of increased sales, and the sixth row tells a tale of constant sales. But what is the “story” about the seventh row? There, the partial icon is so small that we have no idea what product category the icon represents.

Perhaps a more serious issue is those curious chart labels. “Latest twelve months” is ambiguous; it could include months from the current year as well as months from one year ago and therefore may not be an equivalent time period to “two years ago.” But the business was established in 2001, and the claim being made is “best sales year ever,” so why hasn’t management included sales figures for *every* year?

Are the Whitney Wireless managers hiding something, or are they just unaware of the proper use of statistics? Either way, they have failed to properly organize and visualize their data and therefore have failed to communicate a vital aspect of their story.

In subsequent Digital Cases, you will be asked to provide this type of analysis, using the open-ended case questions as your guide. Not all the cases are as straightforward as this example, and some cases include perfectly appropriate applications of statistical methods.

CHAPTER 1 EXCEL GUIDE

EG1.1 DEFINING DATA

Establishing the Variable Type

Microsoft Excel infers the variable type from the data you enter into a column. If Excel discovers a column that contains numbers, for example, it treats the column as a numerical variable. If Excel discovers a column that contains words or alphanumeric entries, it treats the column as a non-numerical (categorical) variable.

This imperfect method works most of the time, especially if you make sure that the categories for your categorical variables are words or phrases such as “yes” and “no.” However, because you cannot explicitly define the variable type, Excel can mistakenly offer or allow you to do nonsensical things such as using a statistical method that is designed for numerical variables on categorical variables. If you must use coded values such as 1, 2, or 3, enter them preceded with an apostrophe, as Excel treats all values that begin with an apostrophe as non-numerical data. (You can check whether a cell entry includes a leading apostrophe by selecting a cell and viewing the contents of the cell in the formula bar.)

EG1.2 MEASUREMENT SCALES for VARIABLES

There are no Excel Guide instructions for this section.

EG1.3 COLLECTING DATA

Recoding Variables

Key Technique To recode a categorical variable, you first copy the original variable’s column of data and then use the find-and-replace function on the copied data. To recode a numerical variable, enter a formula that returns a recoded value in a new column.

Example Using the **DATA worksheet** of the **Recoded workbook**, create the recoded variable UpperLower from the categorical variable Class and create the recoded Variable Dean’s List from the numerical variable GPA.

In-Depth Excel Use the **RECODED worksheet** of the **Recoded workbook** as a model.

The worksheet already contains UpperLower, a recoded version of Class that uses the operational definitions on page 21, and Dean’s List, a recoded version of GPA, in which the value No recodes all GPA values less than 3.3 and Yes recodes all values 3.3 or greater than 3.3. The **RECODED_FORMULAS worksheet** in the same workbook shows how formulas in column I use the IF function to recode GPA as the Dean’s List variable.

These recoded variables were created by first opening to the **DATA worksheet** in the same workbook and then following these steps:

1. Right-click column D (right-click over the shaded “D” at the top of column D) and click **Copy** in the shortcut menu.
2. Right-click column H and click the **first choice** in the **Paste Options** gallery.

3. Enter **UpperLower** in cell H1.

4. Select column H. With column H selected, click **Home** → **Find & Select** → **Replace**.

In the Replace tab of the Find and Replace dialog box:

5. Enter **Senior** as **Find what**, **Upper** as **Replace with**, and then click **Replace All**.
6. Click **OK** to close the dialog box that reports the results of the replacement command.
7. Still in the Find and Replace dialog box, enter **Junior** as **Find what** (replacing **Senior**), and then click **Replace All**.
8. Click **OK** to close the dialog box that reports the results of the replacement command.
9. Still in the Find and Replace dialog box, enter **Sophomore** as **Find what**, **Lower** as **Replace with**, and then click **Replace All**.
10. Click **OK** to close the dialog box that reports the results of the replacement command.
11. Still in the Find and Replace dialog box, enter **Freshman** as **Find what** and then click **Replace All**.
12. Click **OK** to close the dialog box that reports the results of the replacement command.

(This creates the recoded variable UpperLower in column H.)

13. Enter **Dean’s List** in cell I1.

14. Enter the formula **=IF(G2 < 3.3, "No", "Yes")** in cell I2.

15. Copy this formula down the column to the last row that contains student data (row 63).

(This creates the recoded variable Dean’s List in column I.)

The RECODED worksheet uses the **IF** function to recode the numerical variable into two categories (see Appendix Section F.4). Numerical variables can also be recoded into multiple categories by using the **VLOOKUP** function. Read the **SHORT TAKES** for Chapter 1 to learn more about this advanced recoding technique.

EG1.4 TYPES of SAMPLING METHODS

Simple Random Sample

Key Technique Use the **RANDBETWEEN(smallest integer, largest integer)** function to generate a random integer that can then be used to select an item from a frame.

Example Create a simple random sample *with* replacement of size 40 from a population of 800 items.

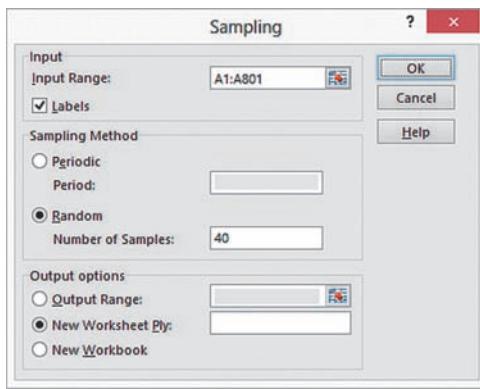
In-Depth Excel Enter a formula that uses this function and then copy the formula down a column for as many rows as is necessary. For example, to create a simple random sample with replacement of size 40 from a population of 800 items, open to a new worksheet. Enter **Sample** in cell A1 and enter the formula **=RANDBETWEEN(1, 800)** in cell A2. Then copy the formula down the column to cell A41.

Excel contains no functions to select a random sample *without* replacement. Such samples are most easily created using an add-in such as PHStat or the Analysis ToolPak, as described in the following paragraphs.

Analysis ToolPak Use **Sampling** to create a random sample *with replacement*.

For the example, open to the worksheet that contains the population of 800 items in column A and that contains a column heading in cell A1. Select **Data** → **Data Analysis**. In the Data Analysis dialog box, select **Sampling** from the **Analysis Tools** list and then click **OK**. In the procedure's dialog box (shown below):

1. Enter A1:A801 as the **Input Range** and check **Labels**.
2. Click **Random** and enter 40 as the **Number of Samples**.
3. Click **New Worksheet Ply** and then click **OK**.



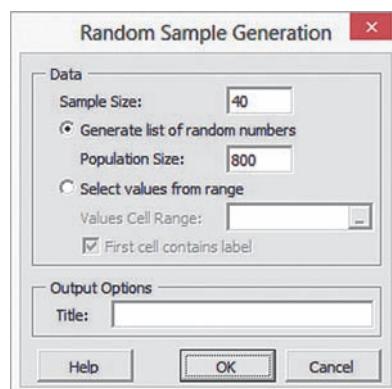
Example Create a simple random sample *without* replacement of size 40 from a population of 800 items.

PHStat Use **Random Sample Generation**.

For the example, select **PHStat** → **Sampling** → **Random Sample Generation**. In the procedure's dialog box (shown below):

1. Enter 40 as the **Sample Size**.
2. Click **Generate list of random numbers** and enter 800 as the **Population Size**.

3. Enter a **Title** and click **OK**.



Unlike most other PHStat results worksheets, the worksheet created contains no formulas.

In-Depth Excel Use the **COMPUTE** worksheet of the **Random** workbook as a template.

The worksheet already contains 40 copies of the formula $=RANDBETWEEN(1, 800)$ in column B. Because the **RANDBETWEEN** function samples *with* replacement as discussed at the start of this section, you may need to add additional copies of the formula in new column B rows until you have 40 unique values.

If your intended sample size is large, you may find it difficult to spot duplicates. Read the **SHORT TAKES** for Chapter 1 to learn more about an advanced technique that uses formulas to detect duplicate values.

CHAPTER 1 MINITAB GUIDE

MG1.1 DEFINING DATA

Establishing the Variable Type

As Section MG.2 mentions on page 11, worksheet columns that contain non-numerical data have names that include “-T” and worksheet columns that contain data that Minitab interprets as either dates or times have names that include “-D.” When Minitab adds a “-T” suffix, it is classifying the column as a categorical, or *text*, variable. When Minitab does not add a suffix, it is classifying the column as a numerical variable. (A column with the “-D” suffix is a *date* variable, a special type of a numerical variable.)

Sometimes, Minitab will misclassify a variable, for example, mistaking a numerical variable for a categorical (text) variable. In such cases, select the column, then select **Data** → **Change Data Type**, and then select one of the choices, for example, **Text to Numeric** for the case of when Minitab has mistaken a numerical variable as a categorical variable.

MG1.2 MEASUREMENT SCALES for VARIABLES

There are no Minitab Guide instructions for this section.

MG1.3 COLLECTING DATA

Recoding Variables

Example Using the DATA worksheet of the Recoded project, create the recoded variable UpperLower from the categorical variable Class (C4-T) and create the recoded variable Dean's List from the numerical variable GPA (C7) to indicate if the GPA value is at least 3.3.

Instructions Use the Replace command to recode a categorical variable. For the example, open to the DATA worksheet of the Recode project and:

1. Select the Class column (C4-T).
2. Select Editor → Replace.

In the Replace in Data Window dialog box:

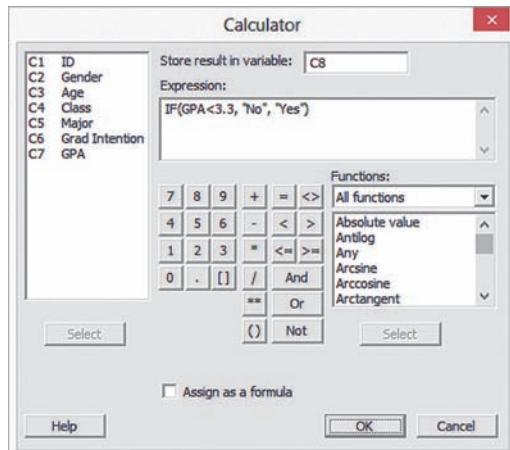
3. Enter Senior as Find what, Upper as Replace with, and then click Replace All.
4. Click OK to close the dialog box that reports the results of the replacement command.
5. Still in the Find and Replace dialog box, enter Junior as Find what (replacing Senior), and then click Replace All.
6. Click OK to close the dialog box that reports the results of the replacement command.
7. Still in the Find and Replace dialog box, enter Sophomore as Find what, Lower as Replace with, and then click Replace All.
8. Click OK to close the dialog box that reports the results of the replacement command.
9. Still in the Find and Replace dialog box, enter Freshman as Find what, and then click Replace All.
10. Click OK to close the dialog box that reports the results of the replacement command.

To create the recoded variable Dean's List from the numerical variable GPA (C7), with the DATA worksheet of the Recode project still open:

1. Enter Dean's List as the name of the empty column C8.
2. Select Calc → Calculator.

In the Calculator dialog box (shown below):

3. Enter C8 in the Store result in variable box.
4. Enter IF(GPA < 3.3, "No", "Yes") in the Expression box.
5. Click OK.



Variables can also be recoded into multiple categories by using the Data → Code command. Read the SHORT TAKES for Chapter 1 to learn more about this advanced recoding technique.

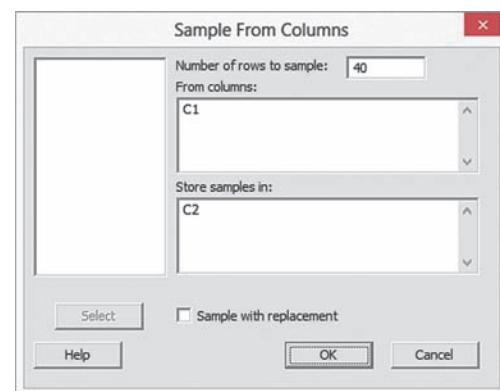
MG1.4 TYPES of SAMPLING METHODS

Simple Random Samples

Example Create a simple random sample with replacement of size 40 from a population of 800 items.

Instructions Use Sample From Columns to create the random sample. For the example, first create the list of 800 employee numbers in column C1. Select Calc → Make Patterned Data → Simple Set of Numbers. In the Simple Set of Numbers dialog box (shown below):

1. Enter C1 in the Store patterned data in box.
2. Enter 1 in the From first value box.
3. Enter 800 in the To last value box.
4. Click OK.

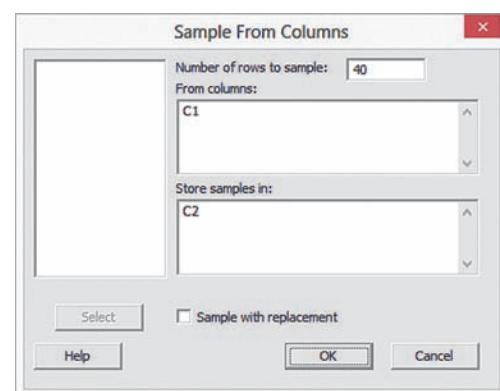


With the worksheet containing the column C1 list still open:

5. Select Calc → Random Data → Sample from Columns.

In the Sample From Columns dialog box (shown below):

6. Enter 40 in the Number of rows to sample box.
7. Enter C1 in the From columns box.
8. Enter C2 in the Store samples in box.
9. Click OK.



CHAPTER

2

Organizing and Visualizing Variables

CONTENTS

2.1 Organizing Categorical Variables

2.2 Organizing Numerical Variables

Classes and Excel Bins

Stacked and Unstacked Data

2.3 Visualizing Categorical Variables

2.4 Visualizing Numerical Variables

2.5 Visualizing Two Numerical Variables

2.6 Organizing Many Categorical Variables

2.7 Challenges in Organizing and Visualizing Variables

Guidelines for Constructing Visualizations

USING STATISTICS: The Choice Is Yours, Revisited

CHAPTER 2 EXCEL GUIDE

CHAPTER 2 MINITAB GUIDE

OBJECTIVES

To construct tables and charts for categorical variables

To construct tables and charts for numerical variables

To learn the principles of properly presenting graphs

To organize and analyze many variables

USING STATISTICS

The Choice Is Yours

Even though he is still in his 20s, Tom Sanchez realizes that he needs to start funding his 401(k) retirement plan now because you can never start too early to save for retirement. Based on research he has already done, Sanchez seeks to invest his money in one or more retirement funds. He decides to contact the *Choice Is Yours* investment service that a business professor had once said was noted for its ethical behavior and fairness toward younger investors.

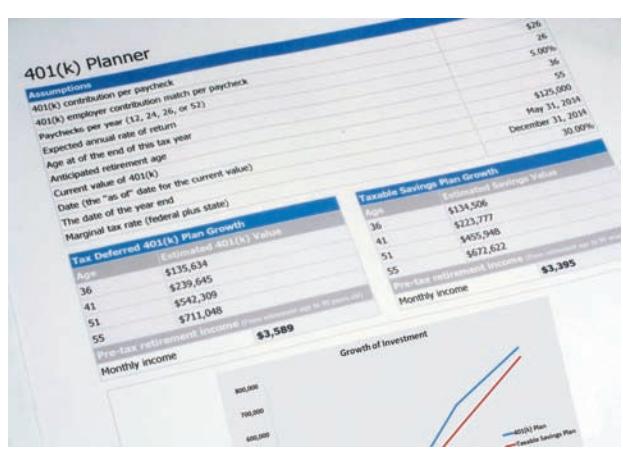
What Sanchez did not know is that *Choice Is Yours* has already been thinking about studying a wide variety of retirement funds, with the business objective of being able to suggest appropriate funds for its younger investors. A company task force has already selected 316 retirement funds that may prove appropriate for younger investors. You have been asked to define, collect, organize, and visualize data about these funds in ways that could assist prospective clients making decisions about the funds in which they will invest. What facts about each fund would you collect to help customers compare and contrast the many funds?

You decide that a good starting point would be to define the variables for key characteristics of each fund, including each fund's past performance. You also decide to define variables such as the amount of assets that a fund manages and whether the goal of a fund is to invest in companies whose earnings are expected to substantially increase in future years (a "growth" fund) or invest in companies whose stock price is undervalued, priced low relative to their earnings potential (a "value" fund).

You collect data from appropriate sources and use the business convention of placing the data for each variable in its own column in a worksheet. As you think more about

your task, you realize that 316 rows of data, one for each fund in the sample, would be hard for prospective clients such as Tom Sanchez to review easily.

Is there something else you can do? Can you organize and present these data to prospective clients in a more helpful and comprehensible manner?




Student Tip

Because of this jumpstart effect, you will find yourself repeating some of the organizing and visualizing methods discussed in this chapter when you study later chapters that focus on methods that help you to analyze the variables.

LEARN MORE

Learn more about retirement funds in general as well as the variables found in the Retirement Funds worksheet in **AllAboutRetirement-Funds.pdf** online section.

Arranging data into columns marks the beginning of the third task of the DCOVA framework, **Organizing** the data collected into tables. While a worksheet full of data columns is a table in the simplest sense, you need to do more for the reasons noted in the scenario.

Designers of the first business computing systems faced a similar problem. Operating under the presumption that the more data shown to decision makers, the better, they created programs that listed all the data collected, one line at a time, in lengthy reports that consumed much paper and that could weigh many pounds. Such reports often failed to facilitate decision making as most decision makers did not have the time to read through a report that could be dozens or hundreds of pages long.

What those decision makers needed was information that summarized the detailed data. Likewise, you need to take the detailed worksheets containing the variables and organize tabular or visual summaries of the data. Many times you do both: You first construct tables that summarize variables and then construct charts and other visual displays based on those tables. In other words, the DCOVA third and fourth tasks, **Organize** and **Visualize** the variables, are often done in tandem or together. When so combined, the **Organize** and **Visualize** tasks can sometimes help jumpstart the **Analysis** task by enabling a decision maker reach preliminary conclusions about data that can be tested during the **Analyze** task.

(Later, in Chapter 17, you will learn that recent advances in computing technology have made practical methods from the interdisciplinary field of business analytics that enables you to combine the organizing and visualizing tasks with the fifth DCOVA task, **Analyze** the data collected to reach conclusions and present results, for even very large or “big” data sets.)

For its examples, this chapter makes extensive use of **Retirement Funds**, the Excel workbook or Minitab worksheet file that contains the sample of 316 funds mentioned in the scenario. This is one of many files that you can download and use with this book as explained in Appendix C.

How to Proceed with This Chapter

Table 2.1 presents the methods used to organize and visualize data that are discussed in this book. This table includes methods for summarizing and describing variables that some instructors prefer to group with the methods of this chapter but which this book discusses in other chapters.

TABLE 2.1

Methods to Organize and Visualize Variables

For Categorical Variables:
Summary table, contingency table (in Section 2.1)
Bar chart, pie chart, Pareto chart, side-by-side bar chart (in Section 2.3)
For Numerical Variables:
Ordered array, frequency distribution, relative frequency distribution, percentage distribution, cumulative percentage distribution (in Section 2.2)
Stem-and-leaf display, histogram, polygon, cumulative percentage polygon (in Section 2.4)
Boxplot (in Section 3.3)
Normal probability plot (in Section 6.3)
Mean, median, mode, quartiles, geometric mean, range, interquartile range, standard deviation, variance, coefficient of variation, skewness, kurtosis (in Sections 3.1, 3.2, and 3.3)
Index numbers (in Section 16.8)
For Two Numerical Variables:
Scatter plot, time-series plot (in Section 2.5)
Sparklines (in Section 17.1)
For Categorical and Numerical Variables Considered Together:
Organizing Many Categorical Variables (in Section 2.6)
Multidimensional contingency tables, PivotTables, gauges, bullet graphs, and treemaps (in Section 17.1)
Cluster analysis (in Section 17.5)
Multidimensional scaling (in Section 17.6)

When you organize the data, you sometimes begin to discover patterns or relationships in the data, as examples in Sections 2.1 and 2.2 illustrate. To better explore and discover patterns and relationships, you can visualize your data by creating various charts and special displays.

Because the methods used to organize and visualize the data collected for categorical variables differ from the methods used to organize and visualize the data collected for numerical variables, this chapter discusses them in separate sections. You will always need to first determine the type of variable, numerical or categorical, you seek to organize and visualize, in order to choose appropriate methods.

This chapter also contains a section on common errors that people make when visualizing variables. When learning methods to visualize variables, you should be aware of such possible errors because of the potential of such errors to mislead and misinform decision makers about the data you have collected.

2.1 Organizing Categorical Variables

You organize categorical variables by tallying the values of a variable by categories and placing the results in tables. Typically, you construct a summary table to organize the data for a single categorical variable and you construct a contingency table to organize the data from two or more categorical variables.

The Summary Table

A **summary table** tallies the values as frequencies or percentages for each category. A summary table helps you see the differences among the categories by displaying the frequency, amount, or percentage of items in a set of categories in a separate column. Table 2.2 presents a summary table that tallies responses to a recent survey that asked young adults about the main reason that they shop online. From this table, stored in **Online Shopping**, you can conclude that 37% shop online mainly for better prices and convenience and that 29% shop online mainly to avoid holiday crowds and hassles.

TABLE 2.2

Main Reason Young Adults Shop Online

Demand	Percentage
Better prices	37%
Avoiding holiday crowds or hassles	29%
Convenience	18%
Better selection	13%
Ships directly	3%

Source: Data extracted and adapted from "Main Reason Young Adults Shop Online?" *USA Today*, December 5, 2012, p. 1A.

EXAMPLE 2.1

Summary Table of Levels of Risk of Retirement Funds

TABLE 2.3

Frequency and Percentage Summary Table of Risk Level for 316 Retirement Funds

The sample of 316 retirement funds for the Choice Is Yours scenario (see page 36) includes the variable risk that has the defined categories low, average, and high. Construct a summary table of the retirement funds, categorized by risk.

SOLUTION From Table 2.3, you can see that about two-thirds of the funds have low risk. About 30% of the funds have average risk. Very few funds have high risk.

Fund Risk Level	Number of Funds	Percentage of Funds
Low	212	67.09%
Average	91	28.80%
High	13	4.11%
Total	316	100.00%

Like worksheet cells, contingency table cells are the intersections of rows and columns, but unlike in a worksheet, both the rows and the columns represent variables. To identify placement, the terms row variable and column variable are often used.

Student Tip

Remember, each joint response gets tallied into only one cell.

The Contingency Table

A **contingency table** cross-tabulates, or tallies jointly, the values of two or more categorical variables, allowing you to study patterns that may exist between the variables. Tallies can be shown as a frequency, a percentage of the overall total, a percentage of the row total, or a percentage of the column total, depending on the type of contingency table you use. Each tally appears in its own **cell**, and there is a cell for each **joint response**, a unique combination of values for the variables being tallied. In the simplest contingency table, one that contains only two categorical variables, the joint responses appear in a table such that the tallies of one variable are located in the rows and the tallies of the other variable are located in the columns.

For the sample of 316 retirement funds for the Choice *Is* Yours scenario, you might create a contingency table to examine whether there is any pattern between the fund type variable and the risk level variable. Because the fund type is one of two possible values (Growth or Value) and the risk level is one of three possible values (Low, Average, or High), there would be six possible joint responses for this table. You could create the table by hand tallying the joint responses for each of the retirement funds in the sample. For example, for the first fund listed in the sample you would add to the tally in the cell that is the intersection of the Growth row and the Low column because the first fund is of type Growth and risk level Low. However, a better choice is to use one of the ways described in the Chapter 2 Excel Guide or Minitab Guide to automate this process.

Table 2.4 presents the completed contingency table after all 316 funds have been tallied. In this table, you can see that there are 143 retirement funds that have the value Growth for the fund type variable and the value Low for the risk level variable and that Growth and Low was the most frequent joint response for the fund type and risk level variables.

TABLE 2.4

Contingency Table Displaying Fund Type and Risk Level

FUND TYPE	RISK LEVEL			Total
	Low	Average	High	
Growth	143	74	10	227
Value	69	17	3	89
Total	212	91	13	316

Contingency tables that display cell values as a percentage of a total can help show patterns between variables. Table 2.5 shows a contingency table that displays values as a percentage of the Table 2.4 overall total (316), Table 2.6 shows a contingency table that displays values as a percentage of the Table 2.4 row totals (227 and 89), and Table 2.7 shows a contingency table that displays values as a percentage of the Table 2.4 column totals (212, 91, and 13).

TABLE 2.5

Contingency Table Displaying Fund Type and Risk Level, Based on Percentage of Overall Total

FUND TYPE	RISK LEVEL			Total
	Low	Average	High	
Growth	45.25%	23.42%	3.16%	71.84%
Value	21.84%	5.38%	0.95%	28.16%
Total	67.09%	28.80%	4.11%	100.00%

TABLE 2.6

Contingency Table Displaying Fund Type and Risk Level, Based on Percentage of Row Total

FUND TYPE	RISK LEVEL			Total
	Low	Average	High	
Growth	63.00%	32.60%	4.41%	100.00%
Value	77.53%	19.10%	3.37%	100.00%
Total	67.09%	28.80%	4.11%	100.00%

TABLE 2.7

Contingency Table
Displaying Fund Type
and Risk Level, Based
on Percentage of
Column Total

FUND TYPE	RISK LEVEL			Total
	Low	Average	High	
Growth	67.45%	81.32%	76.92%	71.84%
Value	32.55%	18.68%	23.08%	28.16%
Total	100.00%	100.00%	100.00%	100.00%

Table 2.5 shows that 71.84% of the funds sampled are growth funds, 28.16% are value funds, and 45.25% are growth funds that have low risk. Table 2.6 shows that 63% of the growth funds have low risk, while 77.53% of the value funds have low risk. Table 2.7 shows that of the funds that have low risk, 67.45% are growth funds. From Tables 2.5–2.7, you see that growth funds are less likely than value funds to have low risk.

Problems for Section 2.1

LEARNING THE BASICS

2.1 A categorical variable has three categories, with the following frequencies of occurrence:

Category	Frequency
A	13
B	28
C	9

- a. Compute the percentage of values in each category.
- b. What conclusions can you reach concerning the categories?

2.2 The following data represent the responses to two questions asked in a survey of 40 college students majoring in business: What is your gender? (M = male; F = female) and What is your major? (A = Accounting; C = Computer Information Systems; M = Marketing):

Gender: M M M F M F F M F M
Major: A C C M A C A A C C

Gender: F M M M M F F M F F
Major: A A A M C M A A A C

Gender: M M M M F M F F M M
Major: C C A A M M C A A A

Gender: F M M M M F M F M M
Major: C C A A A A C C A C

- a. Tally the data into a contingency table where the two rows represent the gender categories and the three columns represent the academic major categories.
- b. Construct contingency tables based on percentages of all 40 student responses, based on row percentages and based on column percentages.

APPLYING THE CONCEPTS

2.3 The following table, stored in **Smartphone Sales**, represents the annual percentage of smartphones sold in 2011, 2012, and 2013 (projected).

Type	2011	2012	2013
Android	47%	60%	58%
iOS	19%	22%	23%
Symbian	19%	5%	1%
RIM	11%	6%	4%
Microsoft	2%	4%	10%
Bada	2%	3%	3%

Source: Data extracted from “How High Can Apple Fly,” *USA Today*, October 5, 2012, p. 1A, 2A.

- a. What conclusions can you reach about the market for smartphones in 2011, 2012, and 2013?
- b. What differences are there in the market for smartphones in 2011, 2012, and 2013?

2.4 The **Edmunds.com** NHTSA Complaints Activity Report contains consumer vehicle complaint submissions by automaker, brand, and category (data extracted from **edmu.in/Ybmpuz**). The following tables, stored in **Automaker1** and **Automaker2**, represent complaints received by automaker and complaints received by category for January 2013.

Automaker	Number
American Honda	169
Chrysler LLC	439
Ford Motor Company	440
General Motors	551
Nissan Motors Corporation	467
Toyota Motor Sales	332
Other	516

- Compute the percentage of complaints for each automaker.
- What conclusions can you reach about the complaints for the different automakers?

Category	Number
Airbags and seatbelts	201
Body and glass	182
Brakes	163
Fuel/emission/exhaust system	240
Interior electronics/hardware	279
Powertrain	1,148
Steering	397
Tires and wheels	71

- Compute the percentage of complaints for each category.
- What conclusions can you reach about the complaints for different categories?

2.5 The 2013 Mortimer Spinks and Computer Weekly Technology Survey reflect the views of technology and digital experts across the United Kingdom (bit.ly/WS4jg3). Respondents were asked, “What is the most important factor influencing the success of a tech start-up?” Assume the following results:

Most Important Factor	Frequency
Leadership	400
Marketing	346
Product	464
Technology	86

- Compute the percentage of values for each factor.
- What conclusions can you reach concerning factors influencing successful tech start-ups?

 **2.6** The following table represents world oil production in millions of barrels a day in the third quarter of 2011:

Region	Oil Production (millions of barrels a day)
Iran	3.53
Saudi Arabia	9.34
Other OPEC countries	22.87
Non-OPEC countries	52.52

Source: International Energy Agency, 2012.

- Compute the percentage of values in each category.
- What conclusions can you reach concerning the production of oil in the third quarter of 2011?

2.7 Visier's Survey of Employers explores how North American organizations are solving the challenges of delivering workforce analytics. Employers were asked what would help them be

successful with human resources metrics and reports. The responses (stored in **Needs**) were as follows:

Needs	Frequency
Easier-to-use analytic tools	127
Faster access to data	41
Improved ability to present and interpret data	123
Improved ability to plan actions	33
Improved ability to predict impacts of my actions	49
Improved relationships to the business line organizations	37

Source: Data extracted from bit.ly/YuWYXc.

- Compute the percentage of values for each response need.
- What conclusions can you reach concerning needs for employer success with human resources metrics and reports?

2.8 A survey of 1,085 adults asked “Do you enjoy shopping for clothing for yourself?” The results indicated that 51% of the females enjoyed shopping for clothing for themselves as compared to 44% of the males. (Data extracted from “Split Decision on Clothes Shopping,” *USA Today*, January 28, 2011, p. 1B.) The sample sizes of males and females were not provided. Suppose that the results were as shown in the following table:

ENJOY SHOPPING	GENDER		
	Male	Female	Total
Yes	238	276	514
No	304	267	571
Total	542	543	1,085

- Construct contingency tables based on total percentages, row percentages, and column percentages.
- What conclusions can you reach from these analyses?

2.9 Each day at a large hospital, hundreds of laboratory tests are performed. The rate of “nonconformances,” tests that were done improperly (and therefore need to be redone), has seemed to be steady, at about 4%. In an effort to get to the root cause of the nonconformances, the director of the lab decided to study the results for a single day. The laboratory tests were subdivided by the shift of workers who performed the lab tests. The results are as follows:

LAB TESTS PERFORMED	SHIFT		
	Day	Evening	Total
Nonconforming	16	24	40
Conforming	654	306	960
Total	670	330	1,000

- Construct contingency tables based on total percentages, row percentages, and column percentages.

- b. Which type of percentage—row, column, or total—do you think is most informative for these data? Explain.
- c. What conclusions concerning the pattern of nonconforming laboratory tests can the laboratory director reach?

2.10 Do social recommendations increase ad effectiveness? A study of online video viewers compared viewers who arrived at an advertising video for a particular brand by following a social media recommendation link to viewers who arrived at the same video by web browsing. Data were collected on whether the viewer could correctly recall the brand being advertised after seeing the video. The results were:

CORRECTLY RECALLED THE BRAND		
ARRIVAL METHOD	Yes	No
Recommendation	407	150
Browsing	193	91

Source: Data extracted from “Social Ad Effectiveness: An Unruly White Paper,” www.unrulymedia.com, January 2012, p. 3.

What do these results tell you about social recommendations?

2.2 Organizing Numerical Variables

You organize numerical variables by creating ordered arrays of one or more variables. This section uses the numerical variable meal cost, which represents the cost of a meal at a restaurant, as the basis for its examples. Because the meal cost data has been collected from a sample of 100 restaurants that can be further categorized by their locations as either “city” or “suburban” restaurants, the variable meal cost raises the common question about how data should be organized in a worksheet when a numerical variable represents data from more than one group. This question is answered at the end of this section, after ordered arrays and distributions are first discussed.

The Ordered Array

An **ordered array** arranges the values of a numerical variable in rank order, from the smallest value to the largest value. An ordered array helps you get a better sense of the range of values in your data and is particularly useful when you have more than a few values. For example, financial analysts reviewing travel and entertainment costs might have the business objective of determining whether meal costs at city restaurants differ from meal costs at suburban restaurants. They collect data from a sample of 50 city restaurants and from a sample of 50 suburban restaurants for the cost of one meal (in \$). Table 2.8A shows the unordered data (stored in **Restaurants**). The lack of ordering prevents you from reaching any quick conclusions about meal costs.

TABLE 2.8A

Meal Cost at 50 City Restaurants and 50 Suburban Restaurants

City Restaurant Meal Costs	
33	26
56	67

Suburban Restaurant Meal Costs	
47	48
41	50

In contrast, Table 2.8B, the ordered array version of the same data, enables you to quickly see that the cost of a meal at the city restaurants is between \$25 and \$80 and that the cost of a meal at the suburban restaurants is between \$26 and \$71.

TABLE 2.8B

Ordered Arrays of Meal Costs at 50 City Restaurants and 50 Suburban Restaurants

City Restaurant Meal Cost	
25	26
48	50

Suburban Restaurant Meal Cost	
26	27
44	44

When your collected data contains a large number of values, reaching conclusions from an ordered array can be difficult. In such cases, creating one of the distributions discussed in the following pages would be a better choice.

The Frequency Distribution

A **frequency distribution** tallies the values of a numerical variable into a set of numerically ordered **classes**. Each class groups a mutually exclusive range of values, called a **class interval**. Each value can be assigned to only one class, and every value must be contained in one of the class intervals.

To create a useful frequency distribution, you must consider how many classes would be appropriate for your data as well as determine a suitable *width* for each class interval. In general, a frequency distribution should have at least 5 and no more than 15 classes because having too few or too many classes provides little new information. To determine the **class interval width** [see Equation (2.1)], you subtract the lowest value from the highest value and divide that result by the number of classes you want the frequency distribution to have.

DETERMINING THE CLASS INTERVAL WIDTH

$$\text{Interval width} = \frac{\text{highest value} - \text{lowest value}}{\text{number of classes}} \quad (2.1)$$

For the city restaurant meal cost data shown in Tables 2.8A and 2.8B, between 5 and 10 classes are acceptable, given the size (50) of that sample. From the city restaurant meal costs ordered array in Table 2.8B, the difference between the highest value of \$80 and the lowest value of \$25 is \$55. Using Equation (2.1), you approximate the class interval width as follows:

$$\frac{55}{10} = 5.5$$

This result suggests that you should choose an interval width of \$5.50. However, your width should always be an amount that simplifies the reading and interpretation of the frequency distribution. In this example, such an amount would be either \$5 or \$10, and you should choose \$10, which creates 7 classes, and not \$5, which creates 13 classes, too many for the sample size of 50.

Because each value can appear in only one class, you must establish proper and clearly defined **class boundaries** for each class. For example, if you chose \$10 as the class interval for the restaurant data, you would need to establish boundaries that would include all the values and simplify the reading and interpretation of the frequency distribution. Because the cost of a city restaurant meal varies from \$25 to \$80, establishing the first class interval as \$20 to less than \$30, the second as \$30 to less than \$40, and so on, until the last class interval is \$80 to less than \$90, would meet the requirements. Table 2.9 contains frequency distributions of the cost per meal for the 50 city restaurants and the 50 suburban restaurants using these class intervals.

TABLE 2.9

Frequency Distributions of the Meal Costs for 50 City Restaurants and 50 Suburban Restaurants

Meal Cost (\$)	City Frequency	Suburban Frequency
20 but less than 30	4	4
30 but less than 40	10	17
40 but less than 50	12	13
50 but less than 60	11	10
60 but less than 70	7	4
70 but less than 80	5	2
80 but less than 90	1	0
Total	50	50

The frequency distribution allows you to reach some preliminary conclusions about the data. For example, Table 2.9 shows that the cost of city restaurant meals is concentrated between \$30 and \$60, as is the cost of suburban restaurant meals. However, many more meals at suburban restaurants cost between \$30 and \$40 than at city restaurants.

Student Tip

The total of the frequency column must always equal the total number of values.

For some charts discussed later in this chapter, class intervals are identified by their **class midpoints**, the values that are halfway between the lower and upper boundaries of each class. For the frequency distributions shown in Table 2.9, the class midpoints are \$25, \$35, \$45, \$55, \$65, \$75, \$85, and \$95. Note that well-chosen class intervals lead to class midpoints that are simple to read and interpret, as in this example.

If the data you have collected do not contain a large number of values, different sets of class intervals can create different impressions of the data. Such perceived changes will diminish as you collect more data. Likewise, choosing different lower and upper class boundaries can also affect impressions.

EXAMPLE 2.2

Frequency Distributions of the One-Year Return Percentages for Growth and Value Funds

As a member of the company task force in The Choice Is Yours scenario (see page 36), you are examining the sample of 316 retirement funds stored in **Retirement Funds**. You want to compare the numerical variable **1YrReturn%**, the one-year percentage return of a fund, for the two subgroups that are defined by the categorical variable **Type** (Growth and Value). You construct separate frequency distributions for the growth funds and the value funds.

SOLUTION The one-year percentage returns for both the growth and value funds are concentrated between 10 and 20 (see Table 2.10).

TABLE 2.10

Frequency Distributions of the One-Year Return Percentage for Growth and Value Funds

One-Year Return Percentage	Growth Frequency	Value Frequency
-15 but less than -10	1	0
-10 but less than -5	0	0
-5 but less than 0	0	0
0 but less than 5	6	2
5 but less than 10	23	12
10 but less than 15	104	29
15 but less than 20	75	37
20 but less than 25	12	8
25 but less than 30	3	1
30 but less than 35	3	0
Total	227	89

In the solution for Example 2.2, the total frequency is different for each group (227 and 89). When such totals differ among the groups being compared, you cannot compare the distributions directly as was done in Table 2.9 because of the chance that the table will be misinterpreted. For example, the frequencies for the class interval “5 but less than 10” for growth and “10 but less than 15” for value look similar—23 and 29—but represent two very different parts of a whole: 23 out of 227 and 29 out of 89, or about 10% and 33%, respectively. When the total frequency differs among the groups being compared, you construct either a relative frequency distribution or a percentage distribution.

Classes and Excel Bins

To make use of Microsoft Excel features that can help you construct a frequency distribution, or any of the other types of distributions discussed in this chapter, you must implement your set of classes as a set of Excel **bins**. While bins and classes are both ranges of values, bins do not have explicitly stated intervals.

You establish bins by creating a column that contains a list of bin numbers arranged in ascending order. Each bin number explicitly states the upper boundary of its bin. Bins' lower boundaries are defined implicitly: A bin's lower boundary is the first value greater than the previous bin number. For the column of bin numbers 4.99, 9.99, and 14.99, the second bin has the explicit upper boundary of 9.99 and has the implicit lower boundary of "values greater than 4.99." Compare this to a class

interval, which defines both the lower and upper boundaries of the class, such as in "0 (lower) but less than 5 (upper)."

Because the first bin number does not have a "previous" bin number, the first bin always has negative infinity as its lower boundary. A common workaround to this problem, used in the examples throughout this book (and in PHStat, too), is to define an extra bin, using a bin number that is slightly lower than the lower boundary value of the first class. This extra bin number, appearing first, will allow the now-second bin number to better approximate the first class, though at the cost of adding an unwanted bin to the results.

In this chapter, Tables 2.9 through 2.13 use class groupings in the form "valueA but less than valueB." You can translate class groupings in this

form into nearly equivalent bins by creating a list of bin numbers that are slightly lower than each *valueB* that appears in the class groupings. For example, the Table 2.10 classes on page 44 could be translated into nearly equivalent bins by using this bin number list: -15.01 (the extra bin number is slightly lower than the first lower boundary value -15), -10.01 (slightly less than -10, -5.01, -0.01, 4.99, 9.99, 14.99, 19.99, 24.99, 29.99, and 34.99).

For class groupings in the form "all values from *valueA* to *valueB*," such as the set 0.0 through 4.9, 5.0 through 9.9, 10.0 through 14.9, and 15.0 through 19.9, you can approximate each class grouping by choosing a bin number slightly more than each *valueB*, as in this list of bin numbers: -0.01 (the extra bin number), 4.99 (slightly more than 4.9), 9.99, 14.99, and 19.99.

The Relative Frequency Distribution and the Percentage Distribution

Relative frequency and percentage distributions present tallies in ways other than as frequencies. A **relative frequency distribution** presents the relative frequency, or proportion, of the total for each group that each class represents. A **percentage distribution** presents the percentage of the total for each group that each class represents. When you compare two or more groups, knowing the proportion (or percentage) of the total for each group is more useful than knowing the frequency for each group, as Table 2.11 demonstrates. Compare this table to Table 2.9 on page 43, which displays frequencies. Table 2.11 organizes the meal cost data in a manner that facilitates comparisons.

TABLE 2.11

Relative Frequency Distributions and Percentage Distributions of the Meal Costs at City and Suburban Restaurants

MEAL COST (\$)	CITY		SUBURBAN	
	Relative Frequency	Percentage	Relative Frequency	Percentage
20 but less than 30	0.08	8.0%	0.08	8.0%
30 but less than 40	0.20	20.0%	0.34	34.0%
40 but less than 50	0.24	24.0%	0.26	26.0%
50 but less than 60	0.22	22.0%	0.20	20.0%
60 but less than 70	0.14	14.0%	0.08	8.0%
70 but less than 80	0.10	10.0%	0.04	4.0%
80 but less than 90	0.02	2.0%	0.00	0.0%
Total	1.00	100.0%	1.00	100.0%

The **proportion**, or **relative frequency**, in each group is equal to the number of *values* in each class divided by the total number of values. The percentage in each group is its proportion multiplied by 100%.

COMPUTING THE PROPORTION OR RELATIVE FREQUENCY

The proportion, or relative frequency, is the number of *values* in each class divided by the total number of values:

$$\text{Proportion} = \text{relative frequency} = \frac{\text{number of values in each class}}{\text{total number of values}} \quad (2.2)$$

If there are 80 values and the frequency in a certain class is 20, the proportion of values in that class is

$$\frac{20}{80} = 0.25$$

and the percentage is

$$0.25 \times 100\% = 25\%$$

You construct a relative frequency distribution by first determining the relative frequency in each class. For example, in Table 2.9 on page 43, there are 50 city restaurants, and the cost per meal at 11 of these restaurants is between \$50 and \$60. Therefore, as shown in Table 2.11, the proportion (or relative frequency) of meals that cost between \$50 and \$60 at city restaurants is

$$\frac{11}{50} = 0.22$$

Student Tip

The total of the relative frequency column must always be 1.00. The total of the percentage column must always be 100.

You construct a percentage distribution by multiplying each proportion (or relative frequency) by 100%. Thus, the proportion of meals at city restaurants that cost between \$50 and \$60 is 11 divided by 50, or 0.22, and the percentage is 22%. Table 2.11 on page 45 presents the relative frequency distribution and percentage distribution of the cost of meals at city and suburban restaurants.

From Table 2.11, you conclude that meal cost is slightly more at city restaurants than at suburban restaurants. You note that 14% of the city restaurant meals cost between \$60 and \$70 as compared to 8% of the suburban restaurant meals and that 20% of the city restaurant meals cost between \$30 and \$40 as compared to 34% of the suburban restaurant meals.

EXAMPLE 2.3

Relative Frequency Distributions and Percentage Distributions of the One-Year Return Percentage for Growth and Value Funds

As a member of the company task force in The Choice Is Yours scenario (see page 36), you want to properly compare the one-year return percentages for the growth and value retirement funds. You construct relative frequency distributions and percentage distributions for these funds.

SOLUTION From Table 2.12, you conclude that the one-year return percentage for the growth funds is lower than the one-year return percentage for the value funds. For example, 45.81% of the growth funds have returns between 10 and 15, while 32.58% of the value funds have returns between 10 and 15. Of the growth funds, 33.04% have returns between 15 and 20 as compared to 41.57% of the value funds.

TABLE 2.12

Relative Frequency Distributions and Percentage Distributions of the One-Year Return Percentage for Growth and Value Funds

ONE-YEAR RETURN PERCENTAGE	GROWTH		VALUE	
	Relative Frequency	Percentage	Relative Frequency	Percentage
–15 but less than –10	0.0044	0.44	0.0000	0.00
–10 but less than –5	0.0000	0.00	0.0000	0.00
–5 but less than 0	0.0000	0.00	0.000	0.00
0 but less than 5	0.0264	2.64	0.0225	2.25
5 but less than 10	0.1013	10.13	0.1348	13.48
10 but less than 15	0.4581	45.81	0.3258	32.58
15 but less than 20	0.3304	33.04	0.4157	41.57
20 but less than 25	0.0529	5.29	0.0899	8.99
25 but less than 30	0.0132	1.32	0.0112	1.12
30 but less than 35	0.0132	1.32	0.0000	0.00
Total	1.0000	100.00	1.0000	100.00

The Cumulative Distribution

The **cumulative percentage distribution** provides a way of presenting information about the percentage of values that are less than a specific amount. You use a percentage distribution as the basis to construct a cumulative percentage distribution.

For example, you might want to know what percentage of the city restaurant meals cost less than \$40 or what percentage cost less than \$50. Starting with the Table 2.11 meal cost percentage distribution for city restaurant meal costs on page 45, you combine the percentages of individual class intervals to form the cumulative percentage distribution. Table 2.13 presents the necessary calculations. From this table, you see that none (0%) of the meals cost less than \$20, 8% of meals cost less than \$30, 28% of meals cost less than \$40 (because 20% of the meals cost between \$30 and \$40), and so on, until all 100% of the meals cost less than \$90.

TABLE 2.13

Developing the Cumulative Percentage Distribution for City Restaurant Meal Costs

From Table 2.11:		Percentage (%) of Meal Costs That Are Less Than the Class Interval Lower Boundary
Class Interval	Percentage (%)	
20 but less than 30	8	0 (there are no meals that cost less than 20)
30 but less than 40	20	$8 = 0 + 8$
40 but less than 50	24	$28 = 8 + 20$
50 but less than 60	22	$52 = 8 + 20 + 24$
60 but less than 70	14	$74 = 8 + 20 + 24 + 22$
70 but less than 80	10	$88 = 8 + 20 + 24 + 22 + 14$
80 but less than 90	2	$98 = 8 + 20 + 24 + 22 + 14 + 10$
90 but less than 100	0	$100 = 8 + 20 + 24 + 22 + 14 + 10 + 2$

Table 2.14 is the cumulative percentage distribution for meal costs that uses cumulative calculations for the city restaurants (shown in Table 2.13) as well as cumulative calculations for the suburban restaurants (which are not shown). The cumulative distribution shows that the cost of suburban restaurant meals is lower than the cost of meals in city restaurants. This distribution shows that 42% of the suburban restaurant meals cost less than \$40 as compared to 28% of the meals at city restaurants; 68% of the suburban restaurant meals cost less than \$50, but only 52% of the city restaurant meals do; and 88% of the suburban restaurant meals cost less than \$60 as compared to 74% of such meals at the city restaurants.

TABLE 2.14

Cumulative Percentage Distributions of the Meal Costs for City and Suburban Restaurants

Meal Cost (\$)	Percentage of City Restaurants Meals That Cost Less Than Indicated Amount	Percentage of Suburban Restaurants Meals That Cost Less Than Indicated Amount
20	0	0
30	8	8
40	28	42
50	52	68
60	74	88
70	88	96
80	98	100
90	100	100
100	100	100

Unlike in other distributions, the rows of a cumulative distribution do not correspond to class intervals. (Recall that class intervals are mutually *exclusive*. The rows of cumulative distributions are not: the next row “down” *includes* all of the rows above it.) To identify a row, you use the lower class boundaries from the class intervals of the percentage distribution as is done in Table 2.14.

EXAMPLE 2.4

Cumulative Percentage Distributions of the One-Year Return Percentage for Growth and Value Funds

As a member of the company task force in The Choice *Is Yours* scenario (see page 36), you want to continue comparing the one-year return percentages for the growth and value retirement funds. You construct cumulative percentage distributions for the growth and value funds.

SOLUTION The cumulative distribution in Table 2.15 indicates that returns are lower for the growth funds than for the value funds. The table shows that 59.03% of the growth funds and 48.31% of the value funds have returns below 15%. The table also reveals that 92.07% of the growth funds have returns below 20 as compared to 89.89% of the value funds.

TABLE 2.15

Cumulative Percentage Distributions of the One-Year Return Percentages for Growth and Value Funds

One-Year Return Percentages	Growth Percentage Less Than Indicated Value	Value Percentage Less Than Indicated Value
-15	0.00	0.00
-10	0.44	0.00
-5	0.44	0.00
0	0.44	0.00
5	3.08	2.25
10	13.22	15.73
15	59.03	48.31
20	92.07	89.89
25	97.36	98.88
30	98.68	100.00
35	100.00	100.00

Stacked and Unstacked Data

When data for a numerical variable have been collected for more than one group, you can enter those data in a worksheet as either *unstacked* or *stacked* data.

In an **unstacked** format, you create separate numerical variables for each group. For example, if you entered the meal cost data used in the examples in this section in unstacked format, you would create two numerical variables—city meal cost and suburban meal cost—enter the top data in Table 2.8A on page 42 as the city meal cost data, and enter the bottom data in Table 2.8A as the suburban meal cost data.

In a **stacked** format, you pair a numerical variable that contains all of the values with a sec-

ond, separate categorical variable that contains values that identify to which group each numerical value belongs. For example, if you entered the meal cost data used in the examples in this section in stacked format, you would create a meal cost numerical variable to hold the 100 meal cost values shown in Table 2.8A and create a second location (categorical) variable that would take the value “City” or “Suburban,” depending upon whether a particular value came from a city or suburban restaurant (the top half or bottom half of Table 2.8A).

Sometimes a particular procedure in a data analysis program will require data to be either stacked (or unstacked), and instructions in the

Excel and Minitab Guides note such requirements when they arise. (Both PHStat and Minitab have commands that allow you to automate the stacking or unstacking of data as discussed in the Excel and Minitab Guides for this chapter.) Otherwise, it makes little difference whether your data are stacked or unstacked. However, if you have multiple numerical variables that represent data from the same set of groups, stacking your data will be the more efficient choice. For this reason, the DATA worksheet in **Restaurants** contains the numerical variable Cost and the categorical variable Location to store the meal cost data for the sample of 100 restaurants as stacked data.

Problems for Section 2.2

LEARNING THE BASICS

- 2.11** Construct an ordered array, given the following data from a sample of $n = 7$ midterm exam scores in accounting:

68 94 63 75 71 88 64

- 2.12** Construct an ordered array, given the following data from a sample of midterm exam scores in marketing:

88 78 78 73 91 78 85

- 2.13** In late 2011 and early 2012, the Universal Health Care Foundation of Connecticut surveyed small business owners across the state that employed 50 or fewer employees. The purpose of the study was to gain insight on the current small business health-care environment. Small business owners were asked if they offered health-care plans to their employees and if so, what portion (%) of the employee monthly health-care premium the business paid. The following frequency distribution was formed to summarize the *portion of premium paid* for 89 small businesses who offer health-care plans to employees:

Portion of Premium Paid (%)	Frequency
less than 1%	2
1% but less than 26%	4
26% but less than 51%	16
51% but less than 76%	21
76% but less than 100%	23
100%	23

Source: Data extracted from “Small Business Owners Need Affordable Health Care: A Small Business Health Care Survey,” Universal Health Care Foundation of Connecticut, April 2012, p. 15.

- a. What percentage of small businesses pays less than 26% of the employee monthly health-care premium?
- b. What percentage of small businesses pays between 26% and 75% of the employee monthly health-care premium?
- c. What percentage of small businesses pays more than 75% of the employee monthly health-care premium?

- 2.14** Data were collected on the Facebook website about the most “liked” fast food brands. The data values (the number of “likes” for each fast food brand) for the brands named ranged from 1.0 million to 29.2 million.

- a. If these values are grouped into six class intervals, indicate the class boundaries.
- b. What class interval width did you choose?
- c. What are the six class midpoints?

APPLYING THE CONCEPTS

- 2.15** The file **BBCost2012** contains the total cost (\$) for four tickets, two beers, four soft drinks, four hot dogs, two game programs, two baseball caps, and parking for one vehicle at each of the 30 Major League Baseball parks during the 2012 season. These costs were

176 337 223 174 233 185 160 225 324 187 196 153
184 217 146 172 300 166 184 224 213 242 172 230
257 152 225 151 224 198

Source: Data extracted from fancostexperience.com/pages/fcx/fci_pdfs8.pdf

- a. Organize these costs as an ordered array.
- b. Construct a frequency distribution and a percentage distribution for these costs.
- c. Around which class grouping, if any, are the costs of attending a baseball game concentrated? Explain.



2.16 The file **Utility** contains the following data about the cost of electricity (in \$) during July 2013 for a random sample of 50 one-bedroom apartments in a large city.

96	171	202	178	147	102	153	197	127	82
157	185	90	116	172	111	148	213	130	165
141	149	206	175	123	128	144	168	109	167
95	163	150	154	130	143	187	166	139	149
108	119	183	151	114	135	191	137	129	158

a. Construct a frequency distribution and a percentage distribution that have class intervals with the upper class boundaries \$99, \$119, and so on.

b. Construct a cumulative percentage distribution.

c. Around what amount does the monthly electricity cost seem to be concentrated?

2.17 How much time do commuters living in or near cities spend waiting in traffic, and how much does this waiting cost them per year? The file **Congestion** contains the time spent waiting in traffic and the yearly cost associated with that waiting for commuters in 31 U.S. cities. (Source: Data extracted from “The High Cost of Congestion,” *Time*, October 17, 2011, p. 18.) For both the time spent waiting in traffic and the yearly cost associated with that waiting data,

a. Construct a frequency distribution and a percentage distribution.

b. Construct a cumulative percentage distribution.

c. What conclusions can you reach concerning the time Americans living in or near cities spend sitting in traffic?

d. What conclusions can you reach concerning the time and cost of waiting in traffic per year?

2.18 How do the average credit scores of people living in different American cities differ? The data in **Credit Scores** is an ordered array of the average credit scores of 143 American cities. (Data extracted from usat.ly/109hZAR.)

a. Construct a frequency distribution and a percentage distribution.

b. Construct a cumulative percentage distribution.

c. What conclusions can you reach concerning the average credit scores of people living in different American cities?

2.19 One operation of a mill is to cut pieces of steel into parts that will later be used as the frame for front seats in an automobile. The steel is cut with a diamond saw and requires the resulting parts to be within ± 0.005 inch of the length specified by the automobile company. Data are collected from a sample of 100 steel parts and stored in **Steel**. The measurement reported is the difference in inches between the actual length of the steel part, as measured by a laser measurement device, and the specified length of the steel part. For example, the first value, -0.002 represents a steel part that is 0.002 inch shorter than the specified length.

a. Construct a frequency distribution and a percentage distribution.

b. Construct a cumulative percentage distribution.

c. Is the steel mill doing a good job meeting the requirements set by the automobile company? Explain.

2.20 A manufacturing company produces steel housings for electrical equipment. The main component part of the housing is a steel trough that is made out of a 14-gauge steel coil. It is produced

using a 250-ton progressive punch press with a wipe-down operation that puts two 90-degree forms in the flat steel to make the trough. The distance from one side of the form to the other is critical because of weatherproofing in outdoor applications. The company requires that the width of the trough be between 8.31 inches and 8.61 inches. The widths of the troughs, in inches, collected from a sample of 49 troughs and stored in **Trough**, are:

8.312	8.343	8.317	8.383	8.348	8.410	8.351	8.373
8.481	8.422	8.476	8.382	8.484	8.403	8.414	8.419
8.385	8.465	8.498	8.447	8.436	8.413	8.489	8.414
8.481	8.415	8.479	8.429	8.458	8.462	8.460	8.444
8.429	8.460	8.412	8.420	8.410	8.405	8.323	8.420
8.396	8.447	8.405	8.439	8.411	8.427	8.420	8.498
					8.409		

a. Construct a frequency distribution and a percentage distribution.

b. Construct a cumulative percentage distribution.

c. What can you conclude about the number of troughs that will meet the company’s requirements of troughs being between 8.31 and 8.61 inches wide?

2.21 The manufacturing company in Problem 2.20 also produces electric insulators. If the insulators break when in use, a short circuit is likely to occur. To test the strength of the insulators, destructive testing in high-powered labs is carried out to determine how much *force* is required to break the insulators. Force is measured by observing how many pounds must be applied to the insulator before it breaks. The force measurements, collected from a sample of 30 insulators and stored in **Force**, are:

1,870	1,728	1,656	1,610	1,634	1,784	1,522	1,696
1,592	1,662	1,866	1,764	1,734	1,662	1,734	1,774
1,550	1,756	1,762	1,866	1,820	1,744	1,788	1,688
1,810	1,752	1,680	1,810	1,652	1,736		

a. Construct a frequency distribution and a percentage distribution.

b. Construct a cumulative percentage distribution.

c. What can you conclude about the strength of the insulators if the company requires a force measurement of at least 1,500 pounds before the insulator breaks?

2.22 The file **Bulbs** contains the life (in hours) of a sample of 40 20-watt compact fluorescent light bulbs produced by Manufacturer A and a sample of 40 20-watt compact fluorescent light bulbs produced by Manufacturer B.

a. Construct a frequency distribution and a percentage distribution for each manufacturer, using the following class interval widths for each distribution:

Manufacturer A: 6,500 but less than 7,500, 7,500 but less than 8,500, and so on.

Manufacturer B: 7,500 but less than 8,500, 8,500 but less than 9,500, and so on.

b. Construct cumulative percentage distributions.

c. Which bulbs have a longer life—those from Manufacturer A or Manufacturer B? Explain.

2.23 The file **Drink** contains the following data for the amount of soft drink (in liters) in a sample of 50 2-liter bottles:

2.109 2.086 2.066 2.075 2.065 2.057 2.052 2.044 2.036 2.038
 2.031 2.029 2.025 2.029 2.023 2.020 2.015 2.014 2.013 2.014
 2.012 2.012 2.012 2.010 2.005 2.003 1.999 1.996 1.997 1.992
 1.994 1.986 1.984 1.981 1.973 1.975 1.971 1.969 1.966 1.967
 1.963 1.957 1.951 1.951 1.947 1.941 1.941 1.938 1.908 1.894

- Construct a cumulative percentage distribution.
- On the basis of the results of (a), does the amount of soft drink filled in the bottles concentrate around specific values?

2.3 Visualizing Categorical Variables

The chart you choose to visualize the data for a single categorical variable depends on whether you seek to emphasize how categories directly compare to each other (bar chart) or how categories form parts of a whole (pie chart), or whether you have data that are concentrated in only a few of your categories (Pareto chart). To visualize the data for two categorical variables, you use a side-by-side bar chart.

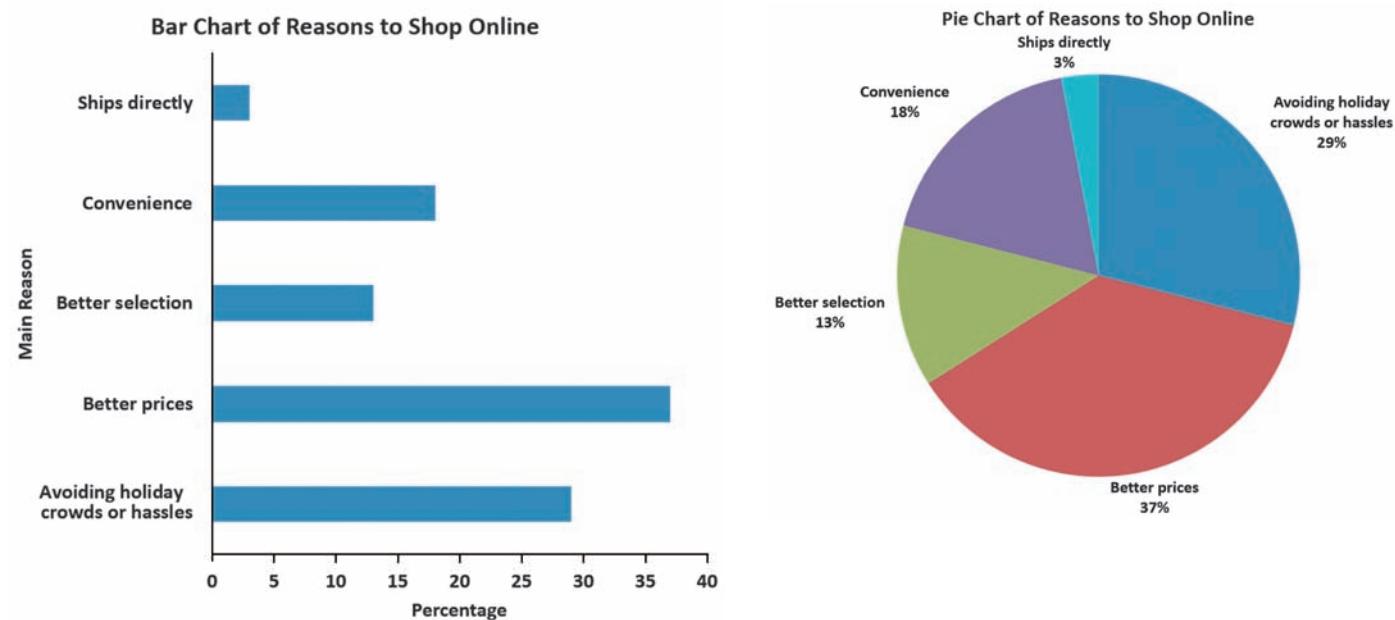
The Bar Chart

A **bar chart** visualizes a categorical variable as a series of bars, with each bar representing the tallies for a single category. In a bar chart, the length of each bar represents either the frequency or percentage of values for a category and each bar is separated by space, called a gap.

The left illustration in Figure 2.1 displays the bar chart for the Table 2.2 summary table on page 38 that tallies responses to a recent survey that asked young adults the main reason they shop online. Reviewing Figure 2.1, you see that respondents are most likely to say because of better prices, followed by avoiding holiday crowds or hassles. Very few respondents mentioned ships directly.

FIGURE 2.1

Excel bar chart (left) and pie chart (right) for reasons for shopping online



EXAMPLE 2.5

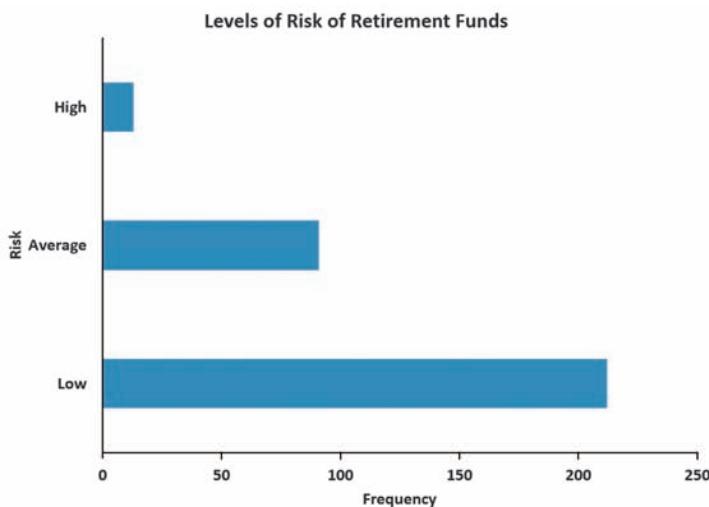
Bar Chart of Levels of Risk of Retirement Funds

FIGURE 2.2

Excel bar chart of the levels of risk of retirement funds

As a member of the company task force in The Choice *Is Yours* scenario (see page 36), you want to first construct a bar chart of the risk of the funds that is based on Table 2.3 on page 38 and then interpret the results.

SOLUTION Reviewing Figure 2.2, you see that low risk is the largest category, followed by average risk. Very few of the funds have high risk.



The Pie Chart

A **pie chart** uses parts of a circle to represent the tallies of each category. The size of each part, or pie slice, varies according to the percentage in each category. For example, in Table 2.2 on page 38, 37% of the respondents stated that they shop online mainly because of better prices. To represent this category as a pie slice, you multiply 37% by the 360 degrees that makes up a circle to get a pie slice that takes up 133.2 degrees of the 360 degrees of the circle, as shown in Figure 2.1 on page 51. From the Figure 2.1 pie chart, you can see that the second largest slice is avoiding holiday crowd and hassles, which contains 29% of the pie.

EXAMPLE 2.6

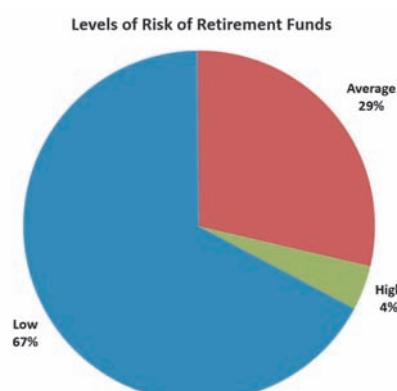
Pie Chart of Levels of Risk of Retirement Funds

FIGURE 2.3

Excel pie chart of the risk of retirement funds

As a member of the company task force in The Choice *Is Yours* scenario (see page 36), you want to visualize the risk level of the funds by constructing a pie chart based on Table 2.3 (see page 38) for the risk variable and then interpret the results.

SOLUTION Reviewing Figure 2.3, you see that more than two-thirds of the funds are low risk, about 30% are average risk, and only about 4% are high risk.



Today, some assert that pie charts should never be used. Others argue that they offer an easily comprehended way to visualize parts of a whole. All commentators agree that variations such as 3D perspective pies and “exploded” pie charts, in which one or more slices are pulled away from the center of a pie, should not be used because of the visual distortions they introduce.

The Pareto Chart

In a **Pareto chart**, the tallies for each category are plotted as vertical bars in descending order, according to their frequencies, and are combined with a cumulative percentage line on the same chart. Pareto charts get their name from the **Pareto principle**, the observation that in many data sets, a few categories of a categorical variable represent the majority of the data, while many other categories represent a relatively small, or trivial, amount of the data.

Pareto charts help you to visually identify the “vital few” categories from the “trivial many” categories so that you can focus on the important categories. Pareto charts are also powerful tools for prioritizing improvement efforts, such as when data are collected that identify defective or nonconforming items.

A Pareto chart presents the bars vertically, along with a cumulative percentage line. The cumulative line is plotted at the midpoint of each category, at a height equal to the cumulative percentage. In order for a Pareto chart to include all categories, even those with few defects, in some situations, you need to include a category labeled Other or Miscellaneous. If you include such a category, you place the bar that represents that category at the far end (to the right) of the X axis.

Using Pareto charts can be an effective way to visualize data for many studies that seek causes for an observed phenomenon. For example, consider a bank study team that wants to enhance the user experience of automated teller machines (ATMs). During this study, the team identifies incomplete ATM transactions as a significant issue and decides to collect data about the causes of such transactions. Using the bank’s own processing systems as a primary data source, causes of incomplete transactions are collected, stored in **ATM Transactions**, and then organized in the Table 2.16 summary table.

The informal “80/20” rule, which states that often 80% of results are from 20% of some thing, such as “80% of the work is done by 20% of the employees,” derives from the Pareto principle.

TABLE 2.16

Summary Table of Causes of Incomplete ATM Transactions

Cause	Frequency	Percentage
ATM malfunctions	32	4.42%
ATM out of cash	28	3.87%
Invalid amount requested	23	3.18%
Lack of funds in account	19	2.62%
Card unreadable	234	32.32%
Warped card jammed	365	50.41%
Wrong keystroke	23	3.18%
Total	724	100.00%

Source: Data extracted from A. Bhalla, “Don’t Misuse the Pareto Principle,” *Six Sigma Forum Magazine*, May 2009, pp. 15–18.

To separate out the “vital few” causes from the “trivial many” causes, the bank study team creates the Table 2.17 summary table, in which the causes of incomplete transactions appear in descending order by frequency, as required for constructing a Pareto chart. The table includes the percentages and cumulative percentages for the reordered causes, which the team then uses to construct the Pareto chart shown in Figure 2.4. In Figure 2.4, the vertical axis on the left represents the percentage due to each cause and the vertical axis on the right represents the cumulative percentage.

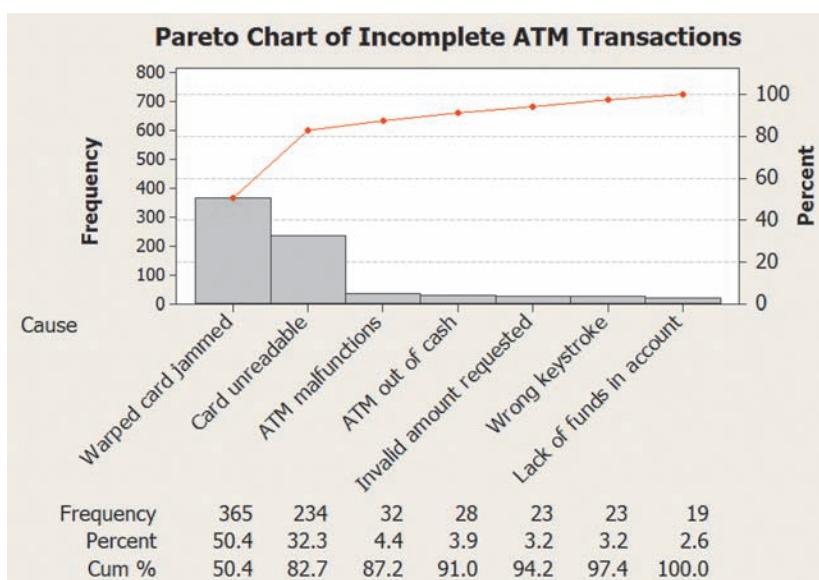
TABLE 2.17

Ordered Summary
Table of Causes of
Incomplete ATM
Transactions

Cause	Frequency	Percentage	Cumulative Percentage
Warped card jammed	365	50.41%	50.41%
Card unreadable	234	32.32%	82.73%
ATM malfunctions	32	4.42%	87.15%
ATM out of cash	28	3.87%	91.02%
Invalid amount requested	23	3.18%	94.20%
Wrong keystroke	23	3.18%	97.38%
Lack of funds in account	19	2.62%	100.00%
Total	724	100.00%	

FIGURE 2.4

Minitab Pareto chart
of incomplete ATM
transactions



Because the categories in a Pareto chart are ordered by decreasing frequency of occurrence, the team can quickly see which causes contribute the most to the problem of incomplete transactions. (Those causes would be the “vital few,” and figuring out ways to avoid such causes would be, presumably, a starting point for improving the user experience of ATMs.) By following the cumulative percentage line in Figure 2.4, you see that the first two causes, warped card jammed (50.44%) and card unreadable (32.3%), account for 82.7% of the incomplete transactions. Attempts to reduce incomplete ATM transactions due to warped or unreadable cards should produce the greatest payoff.

EXAMPLE 2.7

Pareto Chart of the Main Reason for Shopping Online

Construct a Pareto chart from Table 2.2 (see page 38), which summarizes the main reason young adults shop online.

SOLUTION First, create a new table from Table 2.2 in which the categories are ordered by descending frequency and columns for percentages and cumulative percentages for the ordered categories are included (not shown). From that table, create the Pareto chart in Figure 2.5.

From Figure 2.5, you see that better prices and avoiding holiday crowds and hassles accounted for 66% of the responses and better prices, avoiding holiday crowds and hassles, convenience, and better selection accounted for 97% of the responses.

FIGURE 2.5

Excel Pareto chart of the main reason for shopping online

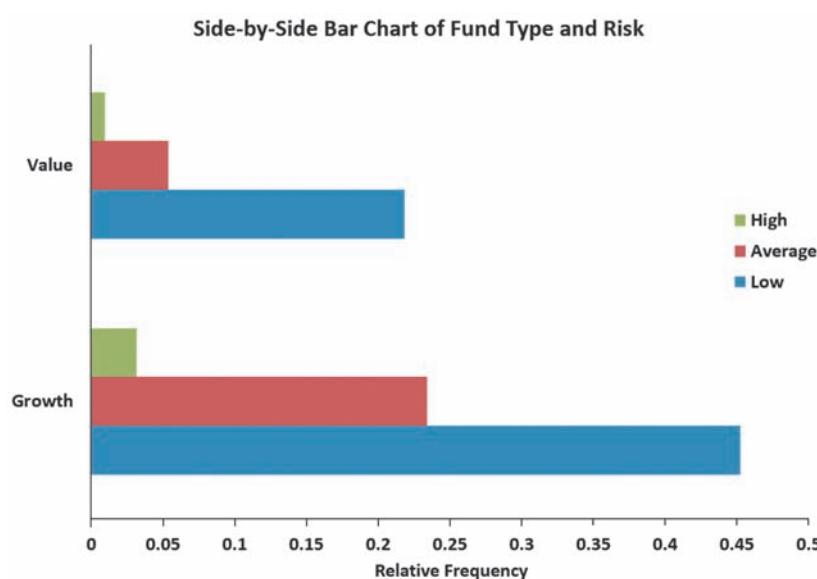


The Side-by-Side Bar Chart

A **side-by-side bar chart** uses sets of bars to show the joint responses from two categorical variables. For example, the Figure 2.6 side-by-side chart visualizes the data for the levels of risk for growth and value funds shown in Table 2.4 on page 39. In Figure 2.6, you see that a substantial portion of the growth funds and the value funds have low risk. However, a larger portion of the growth funds have average risk.

FIGURE 2.6

Side-by-side bar chart of fund type and risk level



Problems for Section 2.3

APPLYING THE CONCEPTS

2.24 An online survey commissioned by Vizu, a Nielsen company, of digital marketing and media professionals on current attitudes and practices regarding paid social media advertising was conducted by Digiday in fall 2012. Advertisers were asked to indicate the primary purpose of their paid social media ads. The survey results were as follows:

Paid Social Media Advertising Objective	Percentage
Primarily branding related, e.g. raising awareness, influencing brand opinions	45%
Primarily direct-response related, e.g. driving product trials or site visits	16%
Mix—more than half is branding	25%
Mix—more than half is direct-response	14%
Source: Data extracted from www.nielsen.com/us/en/reports/2013/the-paid-social-media-advertising-report-2013.html .	

- a. Construct a bar chart, a pie chart, and a Pareto chart.
- b. Which graphical method do you think is best for portraying these data?
- c. What conclusions can you reach concerning the purpose of paid media advertising?

2.25 What do college students do with their time? A survey of 3,000 traditional-age students was taken, with the results as follows:

Activity	Percentage
Attending class/lab	9%
Sleeping	24%
Socializing, recreation, other	51%
Studying	7%
Working, volunteering, student clubs	9%
Source: Data extracted from M. Marklein, "First Two Years of College Wasted?" <i>USA Today</i> , January 18, 2011, p. 3A.	

- a. Construct a bar chart, a pie chart, and a Pareto chart.
- b. Which graphical method do you think is best for portraying these data?
- c. What conclusions can you reach concerning what college students do with their time?

2.26 The Energy Information Administration reported the following sources of electricity in the United States in 2011:

Source of Electricity	Percentage
Coal	42%
Hydro and renewables	13%
Natural gas	25%
Nuclear power	19%
Other	1%
Source: Energy Information Administration, 2011.	

- a. Construct a Pareto chart.
- b. What percentage of power is derived from coal, nuclear power, or natural gas?

- c. Construct a pie chart.

d. For these data, do you prefer using a Pareto chart or a pie chart? Why?

2.27 The **Edmunds.com** NHTSA Complaints Activity Report contains consumer vehicle complaint submissions by automaker, brand, and category (data extracted from edmu.in/Ybmpuz.) The following tables, stored in **Automaker1** and **Automaker2**, represent complaints received by automaker and complaints received by category for January 2013.

Automaker	Number
American Honda	169
Chrysler LLC	439
Ford Motor Company	440
General Motors	551
Nissan Motors Corporation	467
Toyota Motor Sales	332
Other	516

- a. Construct a bar chart and a pie chart for the complaints received by automaker.
- b. Which graphical method do you think is best for portraying these data?

Category	Number
Airbags and seatbelts	201
Body and glass	182
Brakes	63
Fuel/emission/exhaust system	240
Interior electronics/hardware	279
Powertrain	1,148
Steering	397
Tires and wheels	71

- c. Construct a Pareto chart for the categories of complaints.
- d. Discuss the "vital few" and "trivial many" reasons for the categories of complaints.

2.28 The following table indicates the percentage of residential electricity consumption in the United States, in a recent year organized by type of use

Type of Appliance	Percentage
Cooking	4%
Cooling	9%
Electronics	6%
Heating	45%
Lighting	6%
Refrigeration	4%
Water heating	18%
Wet cleaning	3%
Other	5%
Source: Department of Energy	

- Construct a bar chart, a pie chart, and a Pareto chart.
- Which graphical method do you think is best for portraying these data?
- What conclusions can you reach concerning residential electricity consumption in the United States?

2.29 Visier's Survey of Employers explores how North American organizations are solving the challenges of delivering workforce analytics. Employers were asked what would help them be successful with human resources metrics and reports. The responses were as follows (stored in **Needs**):

Needs	Frequency
Easier-to-use analytic tools	127
Faster access to data	41
Improved ability to present and interpret data	123
Improved ability to plan actions	33
Improved ability to predict impacts of my actions	49
Improved relationships to the business line organizations	37

Source: Data extracted from bit.ly/YuWYXc.

- Construct a bar chart and a pie chart.
- What conclusions can you reach concerning needs for employer success with human resource metrics and reports?

2.30 A survey of 1,085 adults asked "Do you enjoy shopping for clothing for yourself?" The results indicated that 51% of the females enjoyed shopping for clothing for themselves as compared to 44% of the males. (Data extracted from "Split Decision on Clothes Shopping," *USA Today*, January 28, 2011, p. 1B.) The sample sizes of males and females were not provided. Suppose that the results were as shown in the following table:

ENJOY SHOPPING FOR CLOTHING	GENDER		
	Male	Female	Total
Yes	238	276	514
No	304	267	571
Total	542	543	1,085

- Construct a side-by-side bar chart of enjoying shopping and gender.
- What conclusions can you reach from this chart?

2.31 Each day at a large hospital, hundreds of laboratory tests are performed. The rate of "nonconformances," tests that were done improperly (and therefore need to be redone), has seemed to be steady, at about 4%. In an effort to get to the root cause of the nonconformances, the director of the lab decided to study the results for a single day. The laboratory tests were subdivided by the shift of workers who performed the lab tests. The results are as follows:

LAB TESTS PERFORMED	SHIFT		Total
	Day	Evening	
Nonconforming	16	24	40
Conforming	654	306	960
Total	670	330	1,000

- Construct a side-by-side bar chart of nonconformances and shift.
- What conclusions concerning the pattern of nonconforming laboratory tests can the laboratory director reach?

2.32 Do social recommendations increase ad effectiveness? A study of online video viewers compared viewers who arrived at an advertising video for a particular brand by following a social media recommendation link to viewers who arrived at the same video by web browsing. Data were collected on whether the viewer could correctly recall the brand being advertised after seeing the video. The results were as follows:

ARRIVAL METHOD	CORRECTLY RECALLED THE BRAND	
	Yes	No
Recommendation	407	150
Browsing	193	91

Source: Data extracted from "Social Ad Effectiveness: An Unruly White Paper," www.unrulymedia.com, January 2012, p. 3.

- Construct a side-by-side bar chart of the arrival method and whether the brand was promptly recalled.
- What do these results tell you about the arrival method and brand recall?

2.4 Visualizing Numerical Variables

You visualize the data for a numerical variable through a variety of techniques that show the distribution of values. These techniques include the stem-and-leaf display, the histogram, the percentage polygon, and the cumulative percentage polygon (ogive), all discussed in this section, as well as the boxplot, which requires descriptive summary measures, as explained in Section 3.3.

The Stem-and-Leaf Display

A **stem-and-leaf display** visualizes data by presenting the data as one or more row-wise *stems* that represent a range of values. In turn, each stem has one or more *leaves* that branch out to the right of their stem and represent the values found in that stem. For stems with more than one leaf, the leaves are arranged in ascending order.

Stem-and-leaf displays allow you to see how the data are distributed and where concentrations of data exist. Leaves typically present the last significant digit of each value, but sometimes you round values. For example, suppose you collect the following meal costs (in \$) for 15 classmates who had lunch at a fast-food restaurant (stored in **FastFood**):

7.42 6.29 5.83 6.50 8.34 9.51 7.10 6.80 5.90 4.89 6.50 5.52 7.90 8.30 9.60

To construct the stem-and-leaf display, you use whole dollar amounts as the stems and round the cents to one decimal place to use as the leaves. For the first value, 7.42, the stem would be 7 and its leaf would be 4. For the second value, 6.29, the stem would be 6 and its leaf 3. The completed stem-and-leaf display for these data is

4	9
5	589
6	3558
7	149
8	33
9	56

Student Tip

If you turn a stem-and-leaf display sideways, the display looks like a histogram.

EXAMPLE 2.8

Stem-and-Leaf Display of the One-Year Return Percentage for the Value Funds

FIGURE 2.7

Minitab stem-and-leaf display of the one-year return percentage for value funds

Using Excel with PHStat will create an equivalent display that contains a different set of stems.

As a member of the company task force in The Choice Is Yours scenario (see page 36), you want to study the past performance of the value funds. One measure of past performance is the numerical variable **1YrReturn%**, the one-year return percentage. Using the data from the 89 value funds, you want to visualize this variable as a stem-and-leaf display.

SOLUTION Figure 2.7 illustrates the stem-and-leaf display of the one-year return percentage for value funds.

Stem-and-Leaf Display: 1YrReturn%_Value

```
Stem-and-leaf of 1YrReturn%_Value  N = 89
Leaf Unit = 1.0
 1    0  1
 1    0
 3    0  45
 7    0  6667
 14   0  8899999
 22   1  00001111
 36   1  22233333333333
(19)  1  444444555555555555
 34   1  66666666667777
 20   1  88889999999
 9    2  00001
 4    2  222
 1    2
 1    2
 1    2
 1    2  8
```

Figure 2.7 allows you to conclude:

- The lowest one-year return was approximately 1.
- The highest one-year return was 28.
- The one-year returns were concentrated between 12 and 19.
- Very few of the one-year returns were above 21.

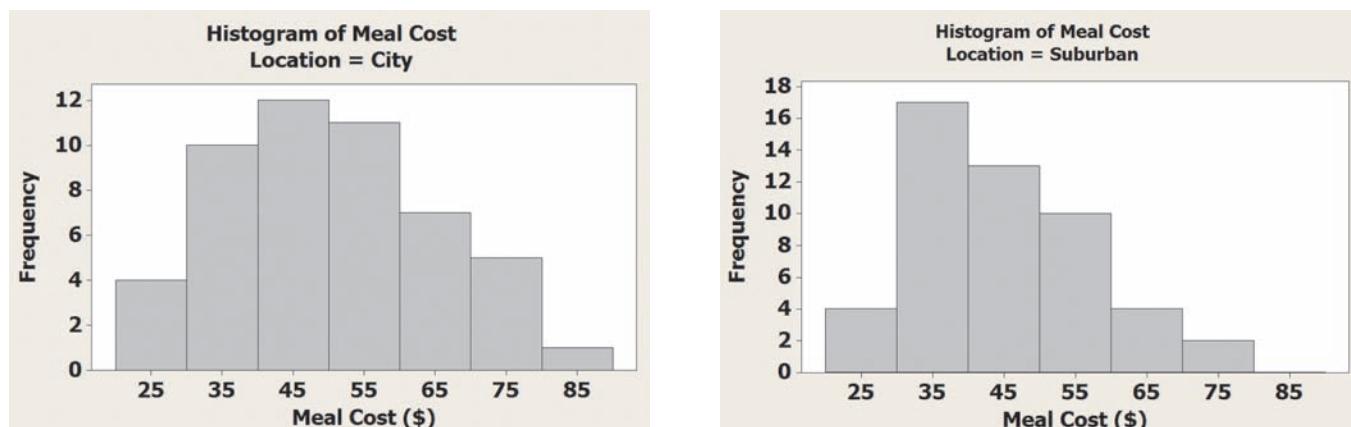
The Histogram

A **histogram** visualizes data as a vertical bar chart in which each bar represents a class interval from a frequency or percentage distribution. In a histogram, you display the numerical variable along the horizontal (X) axis and use the vertical (Y) axis to represent either the frequency or the percentage of values per class interval. There are never any gaps between adjacent bars in a histogram.

Figure 2.8 visualizes the data of Table 2.9 on page 43, meal costs at city and suburban restaurants, as a pair of frequency histograms. The histogram for city restaurants shows that the cost of meals is concentrated between approximately \$30 and \$60. Only one meal at city restaurants cost more than \$80. The histogram for suburban restaurants shows that the cost of meals is also concentrated between \$30 and \$60. However, many more meals at suburban restaurants cost between \$30 and \$40 than at city restaurants. Very few meals at suburban restaurants cost more than \$70.

FIGURE 2.8

Minitab frequency histograms for meal costs at city and suburban restaurants



EXAMPLE 2.9

Histograms of the One-Year Return Percentages for the Growth and Value Funds

As a member of the company task force in The Choice *Is Yours* scenario (see page 36), you seek to compare the past performance of the growth funds and the value funds, using the one-year return percentage variable. Using the data from the sample of 316 funds, you construct histograms for the growth and the value funds to create a visual comparison.

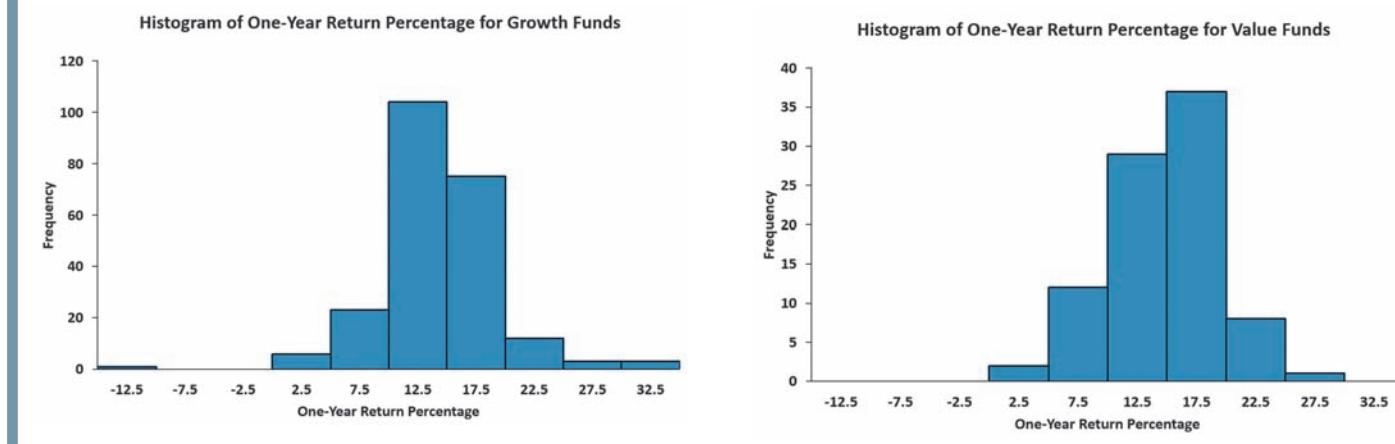
SOLUTION Figure 2.9 displays frequency histograms for the one-year return percentages for the growth and value funds.

Reviewing the histograms in Figure 2.9 leads you to conclude that the returns were lower for the growth funds than for value funds. The return for both the growth funds and the value funds is concentrated between 10 and 20, but the return for the value funds is more concentrated between 15 and 20 while the return for the growth funds is more concentrated between 10 and 15.

(continued)

FIGURE 2.9

Excel frequency histograms for the one-year return percentages for the growth and value funds



The Percentage Polygon

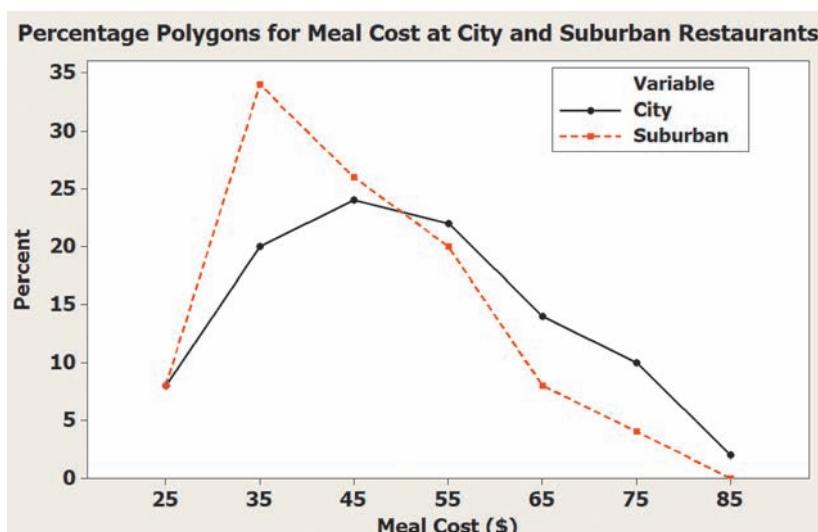
When using a categorical variable to divide the data of a numerical variable into two or more groups, you visualize data by constructing a **percentage polygon**. This chart uses the midpoints of each class interval to represent the data of each class and then plots the midpoints, at their respective class percentages, as points on a line along the X axis. While you can construct two or more histograms, as was done in Figures 2.8 and 2.9, a percentage polygon allows you to make a direct comparison that is easier to interpret. (You cannot, of course, combine two histograms into one chart as bars from the two groups would overlap and obscure data.)

Figure 2.10 displays percentage polygons for the cost of meals at city and suburban restaurants. Compare this figure to the pair of histograms in Figure 2.8 on page 59. Reviewing the polygons in Figure 2.10 allows you to make the same observations as were made when examining Figure 2.8, including the fact that while city restaurant meal costs are both concentrated between \$30 and \$60, suburban restaurants have a much higher concentration between \$30 and \$40. However, unlike the pair of histograms, the polygons allow you to more easily identify which class intervals have similar percentages for the two groups and which do not.

The polygons in Figure 2.10 have points whose values on the X axis represent the midpoint of the class interval. For example, look at the points plotted at $X = 35$ (\$35). The point for meal costs at city restaurants (the lower one) show that 20% of the meals cost between \$30 and \$40, while the point for the meal costs at suburban restaurants (the higher one) shows that 34% of meals at these restaurants cost between \$30 and \$40.

FIGURE 2.10

Minitab percentage polygons of meal costs for city and suburban restaurants



When you construct polygons or histograms, the vertical (Y) axis should include zero to avoid distorting the character of the data. The horizontal (X) axis does not need to show the zero point for the numerical variable, but a major portion of the axis should be devoted to the entire range of values for the variable.

EXAMPLE 2.10

Percentage Polygons of the One-Year Return Percentage for the Growth and Value Funds

FIGURE 2.11

Excel percentage polygons of the one-year return percentages for the growth and value funds

As a member of the company task force in The Choice *Is Yours* scenario (see page 36), you seek to compare the past performance of the growth funds and the value funds using the one-year return percentage variable. Using the data from the sample of 316 funds, you construct percentage polygons for the growth and value funds to create a visual comparison.

SOLUTION Figure 2.11 displays percentage polygons of the one-year return percentage for the growth and value funds.

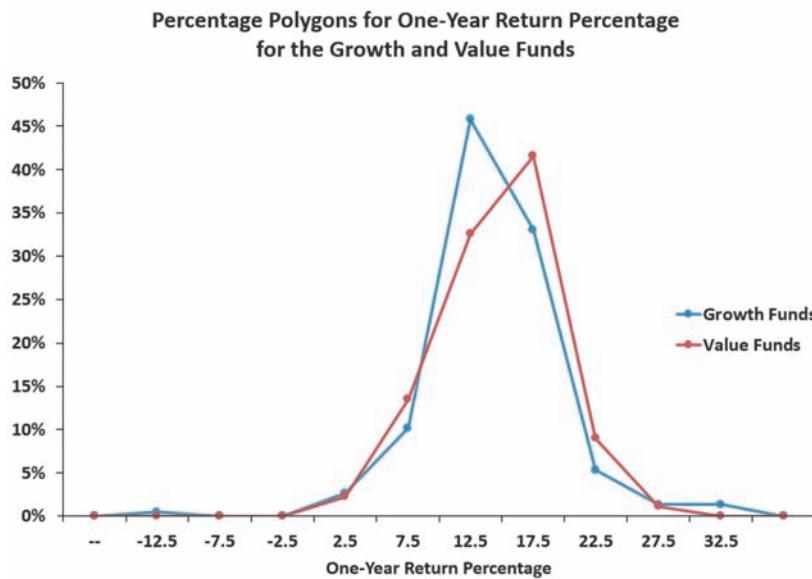


Figure 2.11 shows that the value funds polygon is to the right of the growth funds polygon. This allows you to conclude that the one-year return percentage is higher for value funds than for growth funds. The polygons also show that the return for value funds is concentrated between 15 and 20, and the return for the growth funds is concentrated between 10 and 15.

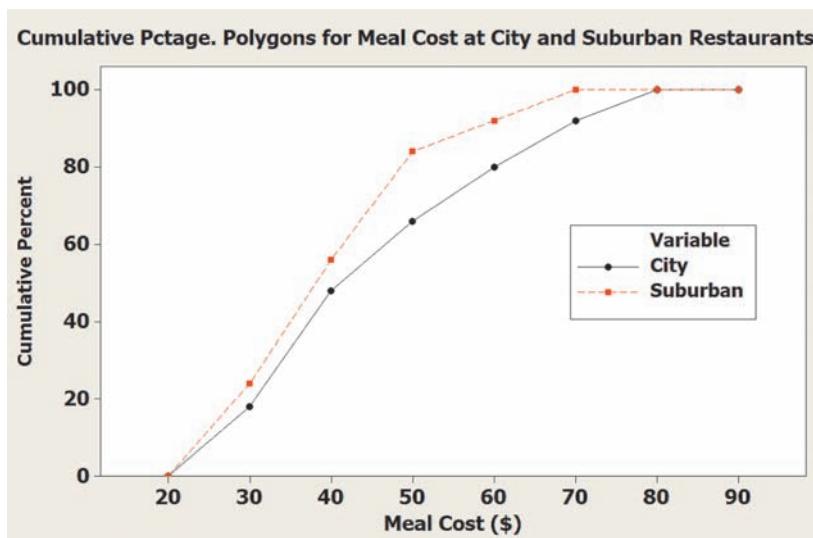
The Cumulative Percentage Polygon (Ogive)

The **cumulative percentage polygon**, or **ogive**, uses the cumulative percentage distribution discussed in Section 2.2 to plot the cumulative percentages along the Y axis. Unlike the percentage polygon, the lower boundary of the class interval for the numerical variable are plotted, at their respective class percentages, as points on a line along the X axis.

Figure 2.12 shows cumulative percentage polygons of meal costs for city and suburban restaurants. In this chart, the lower boundaries of the class intervals (20, 30, 40, etc.) are approximated by the upper boundaries of the previous bins (19.99, 29.99, 39.99, etc.). Reviewing the curves leads you to conclude that the curve of the cost of meals at the city restaurants is located to the right of the curve for the suburban restaurants. This indicates that the city restaurants have fewer meals that cost less than a particular value. For example, 52% of the meals at city restaurants cost less than \$50, as compared to 68% of the meals at suburban restaurants.

FIGURE 2.12

Minitab cumulative percentage polygons of meal costs for city and suburban restaurants

**EXAMPLE 2.11**

Cumulative Percentage Polygons of the One-Year Return Percentages for the Growth and Value Funds

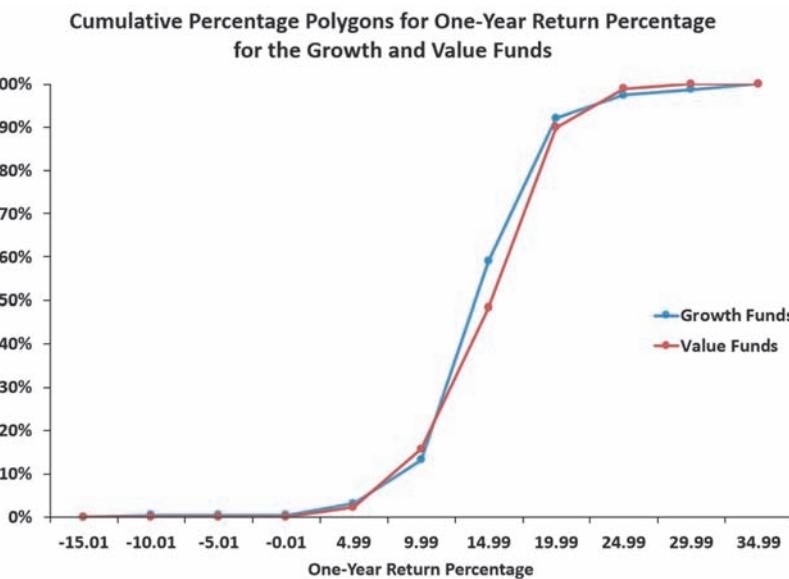
FIGURE 2.13

Excel cumulative percentage polygons of the one-year return percentages for the growth and value funds

In Microsoft Excel, you approximate the lower boundary by using the upper boundary of the previous bin.

As a member of the company task force in The Choice Is Yours scenario (see page 36), you seek to compare the past performance of the growth funds and the value funds using the one-year return percentage variable. Using the data from the sample of 316 funds, you construct cumulative percentage polygons for the growth and the value funds.

SOLUTION Figure 2.13 displays cumulative percentage polygons of the one-year return percentages for the growth and value funds.



The cumulative percentage polygons in Figure 2.13 show that the curve for the one-year return percentage for the growth funds is located slightly to the left of the curve for the value funds. This allows you to conclude that the growth funds have fewer one-year return percentages that are higher than a particular value. For example, 59.03% of the growth funds had one-year return percentages below 15, as compared to 48.31% of the value funds. You can conclude that, in general, the value funds slightly outperformed the growth funds in their one-year returns.

Problems for Section 2.4

LEARNING THE BASICS

2.33 Construct a stem-and-leaf display, given the following data from a sample of midterm exam scores in finance:

54 69 98 93 53 74

2.34 Construct an ordered array, given the following stem-and-leaf display from a sample of $n = 7$ midterm exam scores in information systems:

5	0
6	
7	446
8	19
9	2

APPLYING THE CONCEPTS

2.35 The following is a stem-and-leaf display representing the amount of gasoline purchased, in gallons (with leaves in tenths of gallons), for a sample of 25 cars that use a particular service station on the New Jersey Turnpike:

9	147
10	02238
11	125566777
12	223489
13	02

- a. Construct an ordered array.
- b. Which of these two displays seems to provide more information? Discuss.
- c. What amount of gasoline (in gallons) is most likely to be purchased?
- d. Is there a concentration of the purchase amounts in the center of the distribution?

SELF Test **2.36** The file **BBCost2012** contains the total cost (in \$) for four tickets, two beers, four soft drinks, four hot dogs, two game programs, two baseball caps, and parking for one vehicle at each of the 30 Major League Baseball parks during the 2012 season.

Source: Data extracted from fancostexperience.com/pages/fcx/fci_pdfs/8.pdfs.

- a. Construct a stem-and-leaf display.
- b. Around what value, if any, are the costs of attending a baseball game concentrated? Explain.

2.37 The file **Caffeine** contains the caffeine content (in milligrams per ounce) for a sample of 26 energy drinks:

3.2 1.5 4.6 8.9 7.1 9.0 9.4 31.2 10.0 10.1
9.9 11.5 11.8 11.7 13.8 14.0 16.1 74.5 10.8 26.3
17.7 113.3 32.5 14.0 91.6 127.4

Source: Data extracted from “The Buzz on Energy-Drink Caffeine,” *Consumer Reports*, December 2012.

- a. Construct an ordered array.
- b. Construct a stem-and-leaf display.

- c. Does the ordered array or the stem-and-leaf display provide more information? Discuss.

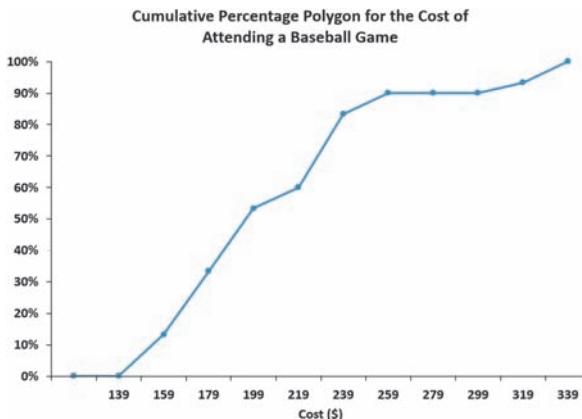
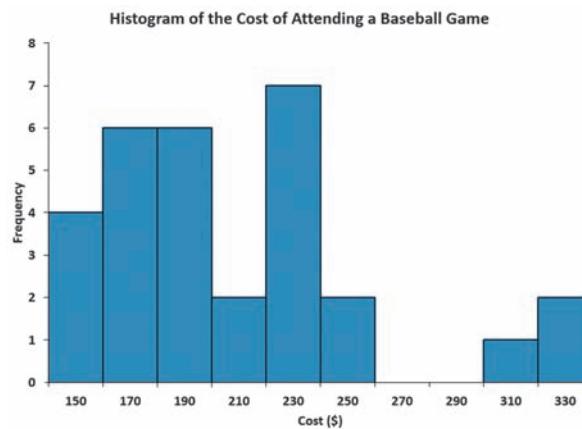
- d. Around what value, if any, is the amount of caffeine in energy drinks concentrated? Explain.

2.38 The file **Utility** contains the following data about the cost of electricity during July 2013 for a random sample of 50 one-bedroom apartments in a large city:

96	171	202	178	147	102	153	197	127	82
157	185	90	116	172	111	148	213	130	165
141	149	206	175	123	128	144	168	109	167
95	163	150	154	130	143	187	166	139	149
108	119	183	151	114	135	191	137	129	158

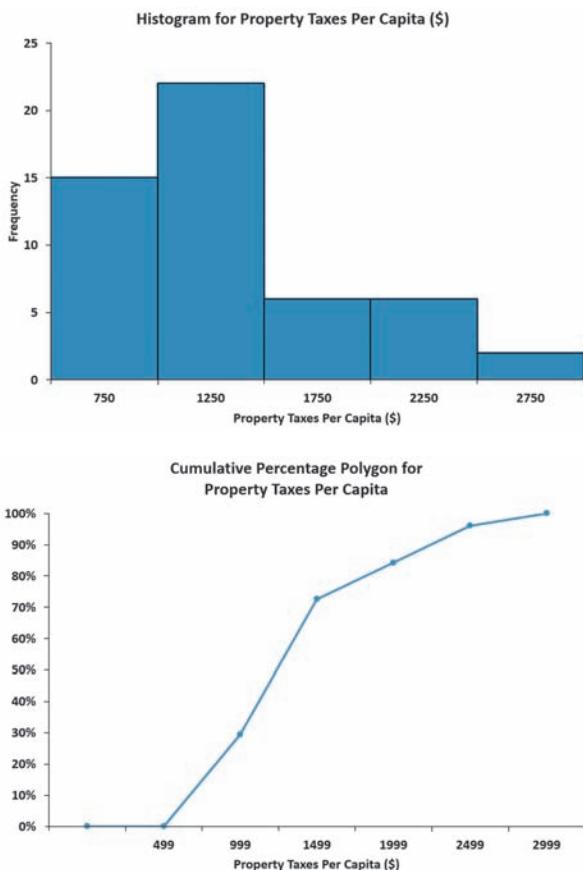
- a. Construct a histogram and a percentage polygon.
- b. Construct a cumulative percentage polygon.
- c. Around what amount does the monthly electricity cost seem to be concentrated?

2.39 As player salaries have increased, the cost of attending baseball games has increased dramatically. The following histogram and cumulative percentage polygon visualizes the total cost (in \$) for four tickets, two beers, four soft drinks, four hot dogs, two game programs, two baseball caps, and parking for one vehicle at each of the 30 Major League Baseball parks during the 2012 season that is stored in **BBCost2012**.



What conclusions can you reach concerning the cost of attending a baseball game at different ballparks?

2.40 The following histogram and cumulative percentage polygon visualize the data about the property taxes per capita(\$) for the 50 states and the District of Columbia, stored in **Property Taxes**.



What conclusions can you reach concerning the property taxes per capita?

2.41 How much time do Americans living in or near cities spend waiting in traffic, and how much does waiting in traffic cost them per year? The data in the file **Congestion** include this cost for 31 cities. (Source: Data extracted from “The High Cost of Congestion,” *Time*, October 17, 2011, p. 18.) For the time Americans living in or near cities spend waiting in traffic and the cost of waiting in traffic per year,

- Construct a percentage histogram.
- Construct a cumulative percentage polygon.
- What conclusions can you reach concerning the time Americans living in or near cities spend waiting in traffic?
- What conclusions can you reach concerning the cost of waiting in traffic per year?

2.42 How do the average credit scores of people living in various cities differ? The file **Credit Scores** contains an ordered array of the average credit scores of 143 American cities. (Data extracted from usat.ly/17a1fA6.)

- Construct a percentage histogram.
- Construct a cumulative percentage polygon.
- What conclusions can you reach concerning the average credit scores of people living in different American cities?

2.43 One operation of a mill is to cut pieces of steel into parts that will later be used as the frame for front seats in an automobile. The steel is cut with a diamond saw and requires the resulting parts to be within ± 0.005 inch of the length specified by the

automobile company. The data are collected from a sample of 100 steel parts and stored in **Steel**. The measurement reported is the difference in inches between the actual length of the steel part, as measured by a laser measurement device, and the specified length of the steel part. For example, the first value, -0.002 represents a steel part that is 0.002 inch shorter than the specified length.

- Construct a percentage histogram.
- Is the steel mill doing a good job meeting the requirements set by the automobile company? Explain.

2.44 A manufacturing company produces steel housings for electrical equipment. The main component part of the housing is a steel trough that is made out of a 14-gauge steel coil. It is produced using a 250-ton progressive punch press with a wipe-down operation that puts two 90-degree forms in the flat steel to make the trough. The distance from one side of the form to the other is critical because of weatherproofing in outdoor applications. The company requires that the width of the trough be between 8.31 inches and 8.61 inches. The widths of the troughs, in inches, collected from a sample of 49 troughs, are stored in **Trough**.

- Construct a percentage histogram and a percentage polygon.
- Plot a cumulative percentage polygon.
- What can you conclude about the number of troughs that will meet the company’s requirements of troughs being between 8.31 and 8.61 inches wide?

2.45 The manufacturing company in Problem 2.44 also produces electric insulators. If the insulators break when in use, a short circuit is likely to occur. To test the strength of the insulators, destructive testing in high-powered labs is carried out to determine how much *force* is required to break the insulators. Force is measured by observing how many pounds must be applied to the insulator before it breaks. The force measurements, collected from a sample of 30 insulators, are stored in **Force**.

- Construct a percentage histogram and a percentage polygon.
- Construct a cumulative percentage polygon.
- What can you conclude about the strengths of the insulators if the company requires a force measurement of at least 1,500 pounds before the insulator breaks?

2.46 The file **Bulbs** contains the life (in hours) of a sample of 40 20-watt compact fluorescent light bulbs produced by Manufacturer A and a sample of 40 20-watt compact fluorescent light bulbs produced by Manufacturer B.

Use the following class interval widths for each distribution:

Manufacturer A: 6,500 but less than 7,500, 7,500 but less than 8,500, and so on.

Manufacturer B: 7,500 but less than 8,500, 8,500 but less than 9,500, and so on.

- Construct percentage histograms on separate graphs and plot the percentage polygons on one graph.
- Plot cumulative percentage polygons on one graph.
- Which manufacturer has bulbs with a longer life—Manufacturer A or Manufacturer B? Explain.

2.47 The data stored in **Drink** represents the amount of soft drink in a sample of 50 2-liter bottles.

- Construct a histogram and a percentage polygon.
- Construct a cumulative percentage polygon.
- On the basis of the results in (a) and (b), does the amount of soft drink filled in the bottles concentrate around specific values?

2.5 Visualizing Two Numerical Variables

Visualizing two numerical variables together can reveal possible relationships between two variables and serve as a basis for applying the methods discussed in Chapters 13 through 17. To visualize two numerical variables, you construct a scatter plot. For the special case in which one of the two variables represents the passage of time, you construct a time-series plot.

The Scatter Plot

A **scatter plot** explores the possible relationship between two numerical variables by plotting the values of one numerical variable on the horizontal, or *X*, axis and the values of a second numerical variable on the vertical, or *Y*, axis. For example, a marketing analyst could study the effectiveness of advertising by comparing advertising expenses and sales revenues of 50 stores by using the *X* axis to represent advertising expenses and the *Y* axis to represent sales revenues.

EXAMPLE 2.12

Scatter Plot for NBA Investment Analysis

Suppose that you are an investment analyst who has been asked to review the valuations of the 30 NBA professional basketball teams. You seek to know if the value of a team reflects its revenues. You collect revenue and valuation data (both in \$millions) for all 30 NBA teams, organize the data as Table 2.18, and store the data in **NBAValues**.

To quickly visualize a possible relationship between team revenues and valuations, you construct a scatter plot as shown in Figure 2.14, in which you plot the revenues on the *X* axis and the value of the team on the *Y* axis.

TABLE 2.18

Revenues and Values for NBA Teams

Team Code	Revenue (\$millions)	Value (\$millions)	Team Code	Revenue (\$millions)	Value (\$millions)	Team Code	Revenue (\$millions)	Value (\$millions)
ATL	99	316	HOU	135	568	OKC	127	475
BOS	143	730	IND	98	383	ORL	126	470
BRK	84	530	LAC	108	430	PHI	107	418
CHA	93	315	LAL	197	1,000	PHX	121	474
CHI	162	800	MEM	96	377	POR	117	457
CLE	128	434	MIA	150	625	SAC	96	525
DAL	137	685	MIL	87	312	SAS	135	527
DEN	110	427	MIN	96	364	TOR	121	405
DET	125	400	NOH	100	340	UTA	111	432
GSW	127	555	NYK	243	1,100	WAS	102	397

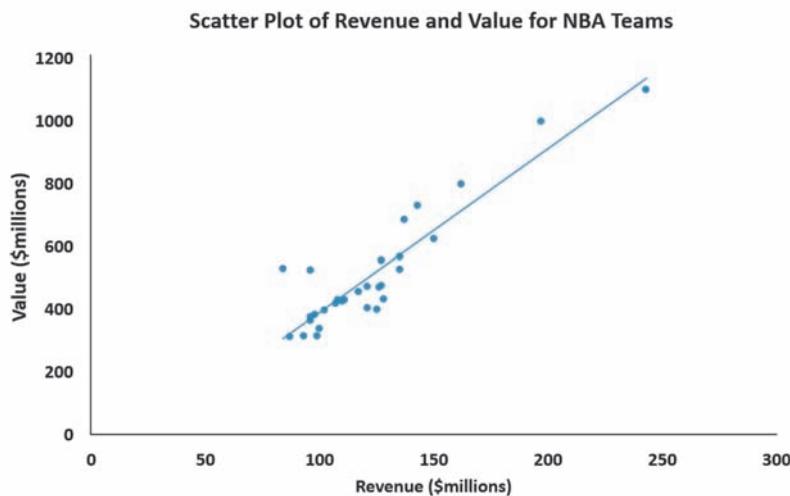
Source: Data extracted from www.forbes.com/nba-valuations.

SOLUTION From Figure 2.14, you see that there appears to be a strong increasing (positive) relationship between revenues and the value of a team. In other words, teams that generate a smaller amount of revenues have a lower value, while teams that generate higher revenues have a higher value. This relationship has been highlighted by the addition of a linear regression prediction line that will be discussed in Chapter 13.

(continued)

FIGURE 2.14

Scatter plot of revenue and value for NBA teams



Other pairs of variables may have a decreasing (negative) relationship in which one variable decreases as the other increases. In other situations, there may be a weak or no relationship between the variables.

LEARN MORE

Read the **SHORT TAKES** for Chapter 2 for an example that illustrates a negative relationship.

The Time-Series Plot

A **time-series plot** plots the values of a numerical variable on the *Y* axis and plots the time period associated with each numerical value on the *X* axis. A time-series plot can help you visualize trends in data that occur over time.

EXAMPLE 2.13
Time-Series Plot for Movie Revenues

As an investment analyst who specializes in the entertainment industry, you are interested in discovering any long-term trends in movie revenues. You collect the annual revenues (in \$billions) for movies released from 1995 to 2012, and organize the data as Table 2.19, and store the data in [Movie Revenues](#).

To see if there is a trend over time, you construct the time-series plot shown in Figure 2.15.

TABLE 2.19

Movie Revenues (in \$billions) from 1995 to 2012

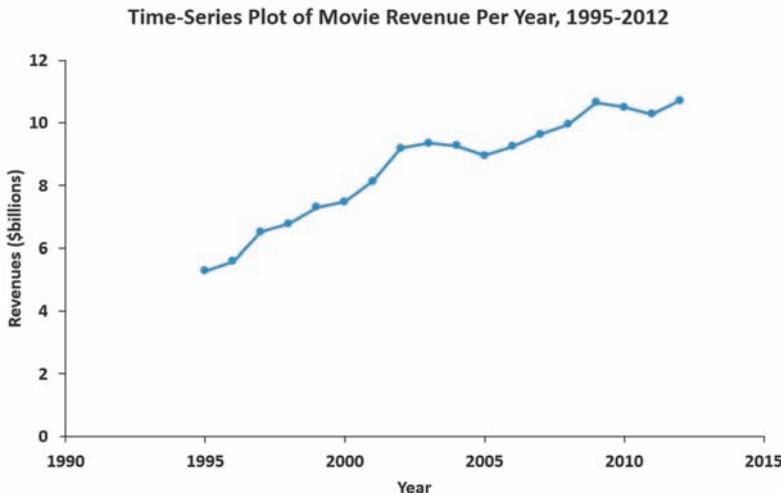
Year	Revenue (\$billions)	Year	Revenue (\$billions)
1995	5.29	2004	9.27
1996	5.59	2005	8.95
1997	6.51	2006	9.25
1998	6.77	2007	9.63
1999	7.30	2008	9.95
2000	7.48	2009	10.65
2001	8.13	2010	10.50
2002	9.19	2011	10.28
2003	9.35	2012	10.71

Source: Data extracted from www.the-numbers.com/market, March 18, 2013.

SOLUTION From Figure 2.15, you see that there was a steady increase in the revenue of movies between 1995 and 2003, a leveling off from 2003 and 2006, followed by a further increase from 2007 to 2009, followed by another leveling off from 2010 to 2012. During that time, the revenue increased from under \$6 billion in 1995 to more than \$10 billion in 2009 to 2012.

FIGURE 2.15

Time-series plot of movie revenue per year from 1995 to 2012



Problems for Section 2.5

LEARNING THE BASICS

2.48 The following is a set of data from a sample of $n = 11$ items:

$$\begin{array}{ccccccccccc} \mathbf{X:} & 7 & 5 & 8 & 3 & 6 & 0 & 2 & 4 & 9 & 5 & 8 \\ \mathbf{Y:} & 1 & 5 & 4 & 9 & 8 & 0 & 6 & 2 & 7 & 5 & 4 \end{array}$$

a. Construct a scatter plot.

b. Is there a relationship between X and Y ? Explain.

2.49 The following is a series of annual sales (in \$millions) over an 11-year period (2002 to 2012):

Year: 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012

Sales: 13.0 17.0 19.0 20.0 20.5 20.5 20.5 20.0 19.0 17.0 13.0

a. Construct a time-series plot.

b. Does there appear to be any change in annual sales over time? Explain.

APPLYING THE CONCEPTS

SELF Test **2.50** Movie companies need to predict the gross receipts of individual movies once a movie has debuted. The following results, stored in **PotterMovies**, are the first weekend gross, the U.S. gross, and the worldwide gross (in \$millions) of the eight Harry Potter movies:

Title	First Weekend (\$millions)	U.S. Gross (\$millions)	Worldwide Gross (\$millions)
<i>Sorcerer's Stone</i>	90.295	317.558	976.458
<i>Chamber of Secrets</i>	88.357	261.988	878.988
<i>Prisoner of Azkaban</i>	93.687	249.539	795.539
<i>Goblet of Fire</i>	102.335	290.013	896.013
<i>Order of the Phoenix</i>	77.108	292.005	938.469
<i>Half-Blood Prince</i>	77.836	301.460	934.601
<i>Deathly Hallows Part I</i>	125.017	295.001	955.417
<i>Deathly Hallows Part II</i>	169.189	381.011	1,328.111

Source: Data extracted from www.the-numbers.com/interactive/comp-HarryPotter.php.

a. Construct a scatter plot with first weekend gross on the X axis and U.S. gross on the Y axis.

b. Construct a scatter plot with first weekend gross on the X axis and worldwide gross on the Y axis.

c. What can you say about the relationship between first weekend gross and U.S. gross and first weekend gross and worldwide gross?

2.51 Data were collected on the typical cost of dining at American-cuisine restaurants within a 1-mile walking distance of a hotel located in a large city. The file **Bundle** contains the typical cost (a per transaction cost in \$) as well as a Bundle score, a measure of overall popularity and customer loyalty, for each of 40 selected restaurants. (Data extracted from www.bundle.com via the link on-msn.com/MnlBxo.)

a. Construct a scatter plot with Bundle score on the X axis and typical cost on the Y axis.

b. What conclusions can you reach about the relationship between Bundle score and typical cost?

2.52 College football is big business, with coaches' pay and revenues in millions of dollars. The file **College Football** contains the coaches' total pay and net revenue for college football at 105 schools. (Data extracted from "College Football Coaches Continue to See Salary Explosion," *USA Today*, November 20, 2012, p. 1C.)

a. Do you think schools with higher net revenues also have higher coaches' pay?

b. Construct a scatter plot with net revenue on the X axis and coaches' pay on the Y axis.

c. Does the scatter plot confirm or contradict your answer to (a)?

2.53 A Pew Research Center survey found that social networking is popular in many nations around the world. The file **Global SocialMedia** contains the level of social media networking (measured as the percentage of individuals polled who use social networking sites) and the GDP at purchasing power parity (PPP) per capita for each of 25 selected countries. (Data extracted from Pew Research Center, "Global Digital Communication: Texting, Social Networking Popular Worldwide," updated February 29, 2012, via the link bit.ly/sNjsmq.)

- Construct a scatterplot with GDP (PPP) per capita on the X axis and social media usage on the Y axis.
- What conclusions can you reach about the relationship between GDP and social media usage?

2.54 How have stocks performed in the past? The following table presents the data stored in **Stock Performance** and shows the performance of a broad measure of stocks (by percentage) for each decade from the 1830s through the 2000s:

Decade	Performance (%)
1830s	2.8
1840s	12.8
1850s	6.6
1860s	12.5
1870s	7.5
1880s	6.0
1890s	5.5
1900s	10.9
1910s	2.2
1920s	13.3
1930s	-2.2
1940s	9.6
1950s	18.2
1960s	8.3
1970s	6.6
1980s	16.6
1990s	17.6
2000s*	-0.5

*Through December 15, 2009.

Source: Data extracted from T. Lauricella, "Investors Hope the '10s Beat the '00s," *The Wall Street Journal*, December 21, 2009, pp. C1, C2.

- Construct a time-series plot of the stock performance from the 1830s to the 2000s.
- Does there appear to be any pattern in the data?

2.55 The data in **NewHomeSales** represent number and median sales price of new single-family houses sold in the United States recorded at the end of each month from January 2000 through December 2012. (Data extracted from www.census.gov, February 24, 2013.)

- Construct a times series plot of new home sales prices.
- What pattern, if any, is present in the data?

2.56 The file **Movie Attendance** contains the yearly movie attendance (in billions) from 2001 through 2012:

Year	Attendance (billions)
2001	1.44
2002	1.58
2003	1.55
2004	1.49
2005	1.40
2006	1.41
2007	1.40
2008	1.39
2009	1.42
2010	1.33
2011	1.30
2012	1.37

Source: Data extracted from the-numbers.com/market.

- Construct a time-series plot for the movie attendance (in billions).
- What pattern, if any, is present in the data?

2.57 The file **Audits** contains the number of audits of corporations with assets of more than \$250 million conducted by the Internal Revenue Service between 2001 and 2012. (Data extracted from www.irs.gov.)

- Construct a time-series plot.
- What pattern, if any, is present in the data?

2.6 Organizing Many Categorical Variables

You construct a **multidimensional contingency table** to tally the responses of three or more categorical variables. In the simplest case of three categorical variables, each cell in the table contains the tallies of the third variable, organized by the subgroups represented by the row and column variables.

Both Excel and Minitab can organize many variables at the same time, but the two programs have different strengths. Using Excel, you can create a **PivotTable**, an interactive table that facilitates exploring multidimensional data. A PivotTable summarizes the variables as a multidimensional table and allows you to interactively change the level of summarization and the arrangement and formatting of the variables. PivotTables also allow you to interactively "slice" your data to summarize subsets of data that meet specified criteria as discussed in Section 17.1.

Methods designed specifically to visualize many categorical variables are beyond the scope of this book to discuss.

Using Minitab, you can also create multidimensional tables, but unlike Excel PivotTables, the Minitab tables are not interactive. However, Minitab, unlike Excel, contains a number of specialized statistical and graphing procedures (beyond the scope of this book to discuss) that can be used to analyze and visualize multidimensional data.

Consider the Table 2.5 contingency table on page 39 that jointly tallies the type and risk variables for the sample of 316 retirement funds as percentages of the overall total. For convenience, this table is shown as a two-dimensional PivotTable in the left illustration of Figure 2.16. This table shows, among other things, that there are many more growth funds of low risk than of average or high risk.

FIGURE 2.16

PivotTables for the retirement funds sample showing percentage of overall total for fund type and showing percentage of overall total risk (left) and for fund type, market cap, and risk (right)

Type	Risk			Grand Total
	Low	Average	High	
Growth	45.25%	23.42%	3.16%	71.84%
Value	21.84%	5.38%	0.95%	28.16%
Grand Total	67.09%	28.80%	4.11%	100.00%

Type	Risk			Grand Total
	Low	Average	High	
Growth	45.25%	23.42%	3.16%	71.84%
Large	28.48%	3.16%	0.95%	32.59%
Mid-Cap	11.71%	11.08%	0.00%	22.78%
Small	5.06%	9.18%	2.22%	16.46%
Value	21.84%	5.38%	0.95%	28.16%
Large	14.24%	1.27%	0.32%	15.82%
Mid-Cap	4.75%	1.27%	0.00%	6.01%
Small	2.85%	2.85%	0.63%	6.33%
Grand Total	67.09%	28.80%	4.11%	100.00%

Adding a third categorical variable, the market cap of the fund, creates the multidimensional contingency table shown at right in Figure 2.16. This new PivotTable reveals the following patterns that cannot be seen in the original Table 2.5 contingency table:

- **For the growth funds, the pattern of risk differs depending on the market cap of the fund.** Large cap funds are most likely to have low risk and are very unlikely to have high risk. Mid-cap funds are equally likely to have low or average risk. Small cap funds are most likely to have average risk and are less likely to have high risk.
- **The value funds show a pattern of risk that is different from the pattern seen in the growth funds.** Mid-cap funds are more likely to have low risk. Almost all of large value funds are low risk, and the small value funds are equally likely to have low or average risk.

Based on these results, the market cap of the mutual fund (small cap, mid-cap, large cap) is an example of a **lurking variable**, a variable that is affecting the results of the other variables. The relationship between the type of fund (growth or value) and the level of risk is clearly affected by the market cap of the mutual fund (small cap, mid-cap, or large cap).

Problems for Section 2.6

APPLYING THE CONCEPTS

2.58 Using the sample of retirement funds stored in **RetirementFunds**:

- Construct a table that tallies type, market cap, and rating.
- What conclusions can you reach concerning differences among the types of retirement funds (growth and value), based on market cap (small, mid-cap, and large) and the rating (one, two, three, four, and five)?

2.59 Using the sample of retirement funds stored in **RetirementFunds**:

- Construct a table that tallies market cap, risk, and rating.
- What conclusions can you reach concerning differences among the types of funds based on market cap (small, mid-cap, and large), risk (low, average, and high), and the rating (one, two, three, four, and five)?

2.60 Using the sample of retirement funds stored in **RetirementFunds**:

- Construct a table that tallies type, risk, and rating.
- What conclusions can you reach concerning differences among the types of retirement funds (growth and value), based on the risk (low, average, and high), and the rating (one, two, three, four, and five)?

2.61 Using the sample of retirement funds stored in **RetirementFunds**:

- Construct a table that tallies type, market cap, risk, and rating.
- What conclusions can you reach concerning differences among the types of funds based on market cap (small, mid-cap, and large), based on type (growth and value), risk (low, average, and high), and rating (one, two, three, four, and five)?

2.7 Challenges in Organizing and Visualizing Variables

As noted throughout this chapter, organizing and visualizing variables can provide useful summaries that can jumpstart the analysis of the variables. However, you must be mindful of the limits of the information technology being used to collect, store, and analyze data as well as the limits of others to be able to perceive and comprehend your results. Many people make a mistake of being overly worried about the former limits—over which, in a typical business environment, they have no control—and forgetting or being naïve about the presentation issues that are often much more critical. You can sometimes easily create summaries and visualizations that obscure the data or create false impressions of the data that lead to misleading or unproductive analysis. The challenge in organizing and visualizing variables is to avoid these complications.

Obscuring Data

Management specialists have long known that *information overload*, presenting too many details, can obscure data and hamper decision making (see reference 2). Figure 2.17 presents an expanded version of the multidimensional contingency table shown in Figure 2.16 on page 69. This table, broken up into two parts by Minitab, illustrates that too many variables as well as data poorly formatted and presented can obscure the data.

FIGURE 2.17

Expanded multidimensional contingency table for the retirement funds sample showing percentage of overall total for fund type, market cap, risk, and star rating

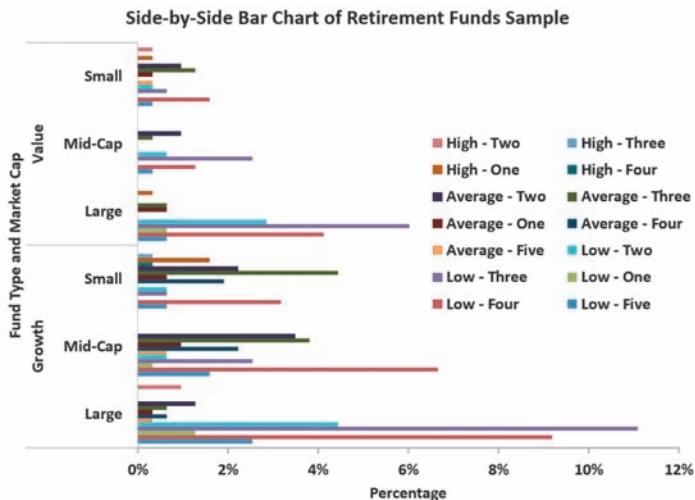
Even though Figure 2.17 uses an example constructed by Minitab, the principle being illustrated holds for Excel as well. The equivalent Excel PivotTable is even more obscuring than the Figure 2.17 table!

Tabulated statistics: Type, Market Cap, Risk, Star Rating									
Rows: Type / Market Cap			Columns: Risk / Star Rating						
	Average								
	Five	Four	One	Three	Two	Five	High	Four	One
Growth									
Large	0.316	0.633	0.316	0.633	1.266	0.000	0.000	0.000	
Mid-Cap	0.633	2.215	0.949	3.797	3.481	0.000	0.000	0.000	
Small	0.000	1.899	0.633	4.430	2.215	0.000	0.316	1.582	
Value									
Large	0.000	0.000	0.633	0.633	0.000	0.000	0.000	0.316	
Mid-Cap	0.000	0.000	0.000	0.316	0.949	0.000	0.000	0.000	
Small	0.316	0.000	0.316	1.266	0.949	0.000	0.000	0.316	
All	All	1.266	4.747	2.848	11.076	8.861	0.000	0.316	2.215
							Low	All	All
							One	Three	Two
Growth									
Large	0.000	0.949	2.532	9.177	1.266	11.076	4.430	32.595	
Mid-Cap	0.000	0.000	1.582	6.646	0.316	2.532	0.633	22.785	
Small	0.316	0.000	0.633	3.165	0.000	0.633	0.633	16.456	
Value									
Large	0.000	0.000	0.633	4.114	0.633	6.013	2.848	15.823	
Mid-Cap	0.000	0.000	0.316	1.266	0.000	2.532	0.633	6.013	
Small	0.000	0.316	0.316	1.582	0.000	0.633	0.316	6.329	
All	All	0.316	1.266	6.013	25.949	2.215	23.418	9.494	100.000
Cell Contents: % of Total									

Visualizations can also be subject to information overload. Figure 2.18 presents a side-by-side bar chart that is based on the obscured data of Figure 2.17 and is typical of charts that sometimes get constructed when using large or complex sets of data, including the “big data” discussed in Chapter 17. As a bar chart, this visualization can highlight certain characteristics of the sample data, consistent with the discussion earlier in the chapter. (For example, when you examine Figure 2.18, you can notice more quickly than you would when examining Figure 2.17 that there are more large-cap retirement funds with low risk and a three-star rating than any other combination of risk and star rating.) However, other details are less obvious, and an overly complex legend poses its own problems even for people who do not suffer from color perception problems.

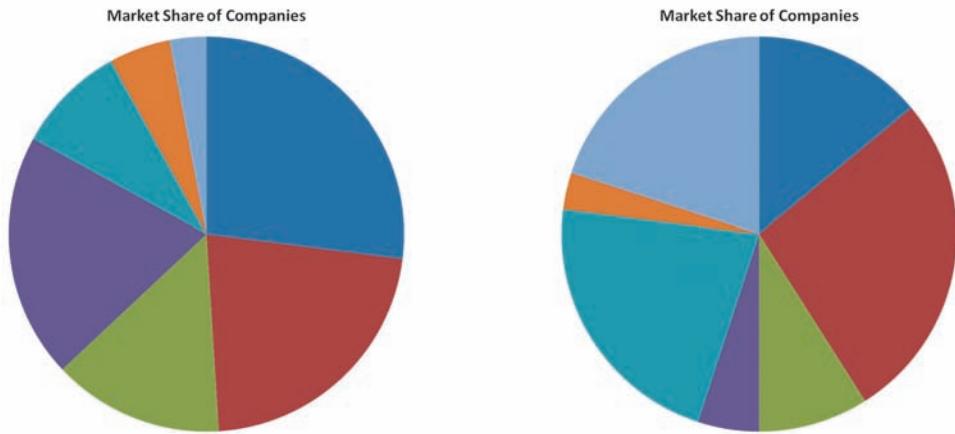
FIGURE 2.18

Side-by-side bar chart for the retirement funds sample showing percentage of overall total for fund type, market cap, risk, and star rating

**FIGURE 2.19**

Market shares of companies in "two" industries

If you are having a hard time believing that the dark blue slice in the left pie chart is equal to the dark red slice in the right pie, open to the *TwoPies* worksheet in the Challenging workbook and verify the underlying data.



Creating False Impressions

As you organize and visualize variables, you must be careful not to create false impressions that could affect preliminary conclusions about the data. Selective summarizations and improperly constructed visualizations often create false impressions.

A *selective summarization* is the presentation of only part of the data that have been collected. Frequently, selective summarization occurs when data collected over a long period of times are summarized as percentage changes for a shorter period. This type of summarization undoes the help that a time-series plot (see page 66) can provide in visualizing trends in data that occur over time. For example, Table 2.20A presents the one-year difference in sales of five auto industry companies for the month of April. That selective summarization tells a different story, particularly for company G, than does Table 2.20B, which shows the year to year differences for three consecutive years.

Improperly constructed charts can also create false impressions. Take a second look at Figure 2.19 above. Because of their relative positions and colorings, many people will perceive the dark blue company in the left chart to have a smaller market share than the dark red company in the right chart even though both pie slices each represent 27%, as can be seen in **False Impressions**. Differently scaled axes in charts visualizing the same data and a *Y* axis that either does not begin at the origin or is a "broken" axis that is missing intermediate values are other common ways to create false impressions.

TABLE 2.20

(A) One-year percentage change in year-to-year sales for the month of April
 (B) Percentage change for three consecutive years

Table 2.20 data is based on sales trends of U.S. automakers during the period 2007–2010 that included the 2008 economic downturn. A times-series plot of these data for a larger period of time would allow you to see trends not evident in Table 2.20B.

Company	Change from Prior Year	Company	Change from Prior Year		
			Year 1	Year 2	Year 3
A	+7.2	A	-22.6	-33.2	+7.2
B	+24.4	B	-4.5	-41.9	+24.4
C	+24.9	C	-18.5	-31.5	+24.9
D	+24.8	D	-29.4	-48.1	+24.8
E	+12.5	E	-1.9	-25.3	+12.5
F	+35.1	F	-1.6	-37.8	+35.1
G	+29.7	G	+7.4	-13.6	+29.7

Chartjunk

Seeking to construct a visualization that can more effectively convey an important point, some people add decorative elements to enhance or replace the simple bar and line shapes of the visualizations discussed in this chapter. While judicious use of such elements may aid in the memorability of a chart (see reference 1), most often such elements either obscure the data or, worse, create a false impression of the data. Such elements are called **chartjunk**.

Figure 2.20 visualizes Australian wine exports to the United States for four years. The chartjunk version on the left uses wine glasses in a histogram-like display in lieu of a proper time-series plot, such as the one shown on the right. Because the years between measurements are not equally spaced, the four wine glasses create a false impression about the ever increasing trend in wine exports. The wine glasses also distort the data by using an object with a three-dimensional volume. (While the height of wine in the 1997 glass is a bit more than 6 times the height of the 1989 glass, the volume of that filled 1997 wine glass would be much more than the almost empty 1989 glass.)

FIGURE 2.20

Two visualizations of Australian wine exports to the United States, in millions of gallons

Chartjunk adapted from S. Watterson, “Liquid Gold—Australians Are Changing the World of Wine. Even the French Seem Grateful,” Time, November 22, 1999, p. 68.



Figure 2.21 presents another visual used in the same magazine article. In the chartjunk version, the grape vine with its leaves and bunch of grapes convey no useful information while creating false impressions about the amount of acreage and the passage of time that obscures the trend easily seen in the time-series plot.

FIGURE 2.21

Two visualizations of the amount of land planted with grapes for the wine industry

Chartjunk adapted from S. Watterson, "Liquid Gold—Australians Are Changing the World of Wine. Even the French Seem Grateful," *Time*, November 22, 1999, pp. 68–69.

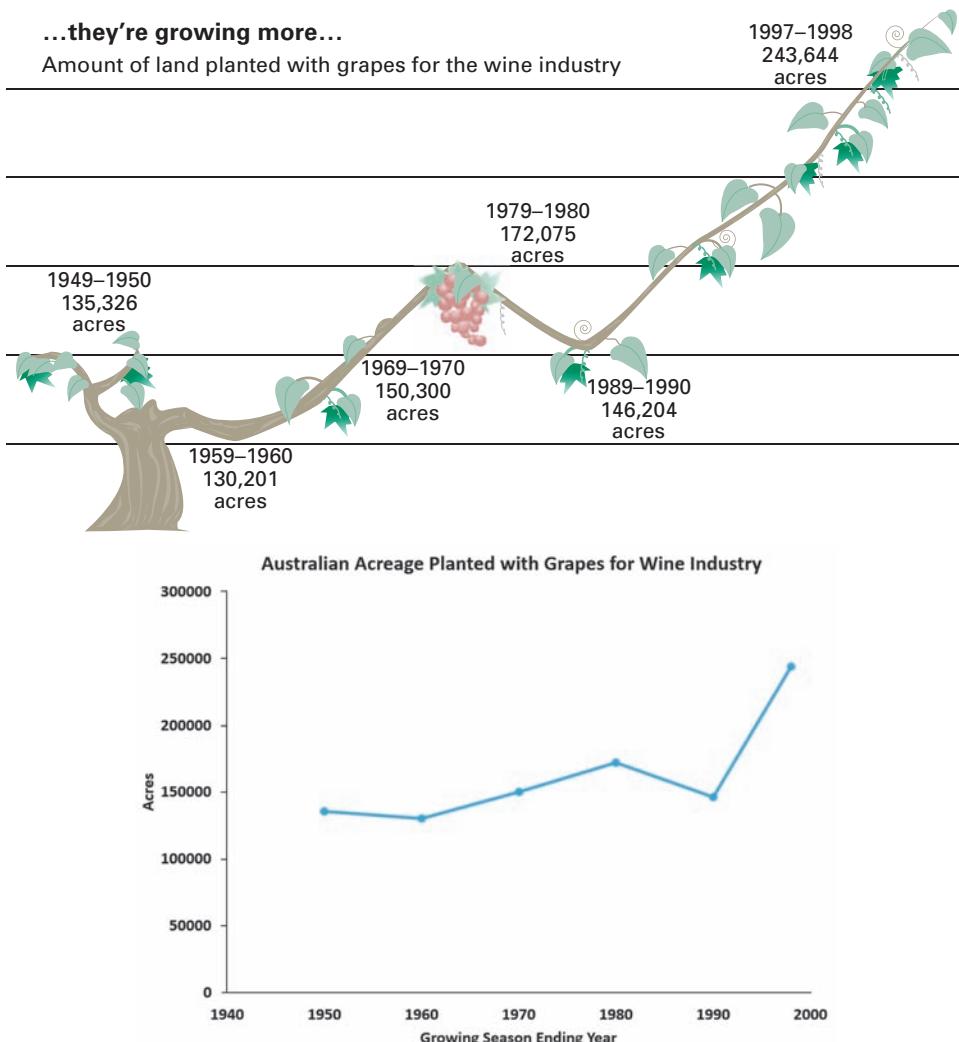
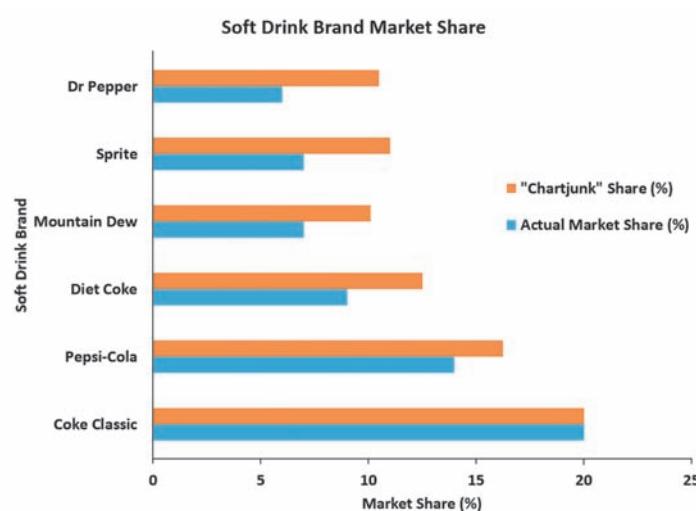
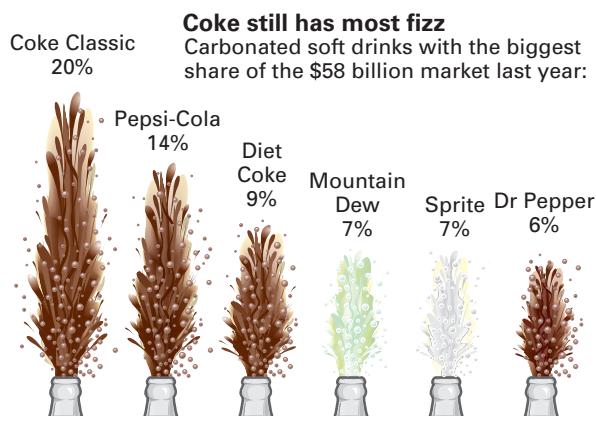


Figure 2.22 visualizes the market share for selected soft drink brands. The chartjunk version fails to convey any more information than a simple bar or pie chart would, and the soft drink bottle tops included in the chart obscure and distort the data. The side-by-side bar chart at right shows how the market shares as represented by the height of the “fizzy” elements overstate the actual market shares of the five lesser brands, using the height of the first bottle and fizz to represent 20%.

FIGURE 2.22

Two visualizations of market share of soft drinks

Chartjunk adapted from Anne B. Carey and Sam Ward, "Coke Still Has Most Fizz," *USA Today*, May 10, 2000, p. 1B.



Guidelines for Constructing Visualizations

To avoid distortions and to create a visualization that best conveys the data, use the following guidelines:

- Use the simplest possible visualization
- Include a title
- Label all axes
- Include a scale for each axis if the chart contains axes
- Begin the scale for a vertical axis at zero
- Use a constant scale

- Avoid 3D effects
- Avoid chartjunk

The grapevine portion of Figure 2.21 on page 73 violates a number of these guidelines besides not avoiding the use of chartjunk. There are no axes present, and there is no clear zero point on the vertical axis. The 1949–1950 acreage, 135,326, is plotted above the higher 1969–1970 acreage, 150,300. Both the horizontal axis and the vertical axis do not contain a constant scale, and neither axis is labeled.

When using Microsoft Excel, beware of such types of distortions. Excel can construct charts in which the vertical axis does not begin at zero and may tempt you to restyle simple charts in an inappropriate manner or may tempt you to use uncommon chart choices such as doughnut, radar, surface, bubble, cone, and pyramid charts. You should resist these temptations as they will often result in a visualization that obscures the data or creates a false impression or both.

Problems for Section 2.7

APPLYING THE CONCEPTS

2.62 (Student Project) Bring to class a chart from a website, newspaper, or magazine published recently that you believe to be a poorly drawn representation of a numerical variable. Be prepared to submit the chart to the instructor with comments about why you believe it is inappropriate. Do you believe that the intent of the chart is to purposely mislead the reader? Also, be prepared to present and comment on this in class.

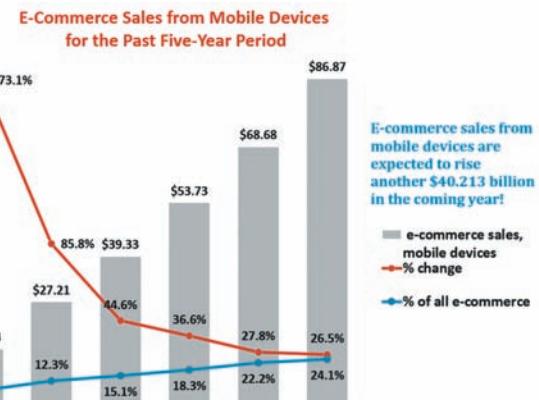
2.63 (Student Project) Bring to class a chart from a website, newspaper, or magazine published this month that you believe to be a poorly drawn representation of a categorical variable. Be prepared to submit the chart to the instructor with comments about why you consider it inappropriate. Do you believe that the intent of the chart is to purposely mislead the reader? Also, be prepared to present and comment on this in class.

2.64 (Student Project) The Data and Story Library (DASL) is an online library of data files and stories that illustrate the use of basic statistical methods. Go to lib.stat.cmu.edu/index.php, click DASL, and explore some of the various graphical displays.

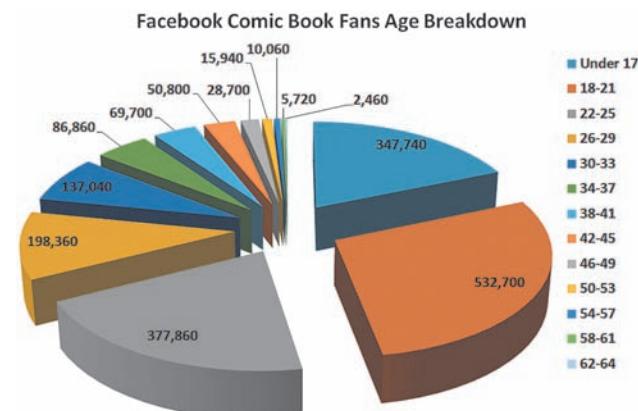
- Select a graphical display that you think does a good job revealing what the data convey. Discuss why you think it is a good graphical display.
- Select a graphical display that you think needs a lot of improvement. Discuss why you think that it is a poorly constructed graphical display.

2.65 Examine the visualization at the top of the next column, adapted from one that appeared in a post in a digital marketing blog.

- Describe at least one good feature of this visual display.
- Describe at least one bad feature of this visual display.
- Redraw the graph, using the guidelines above.



2.66 Examine the following visualization, adapted from one that appeared in the post “Who Are the Comic Book Fans on Facebook?” on February 2, 2013, as reported by graphicspolicy.com.



- Describe at least one good feature of this visual display.
- Describe at least one bad feature of this visual display.
- Redraw the graph, using the guidelines given on page 74.

2.67 Examine the following visualization, adapted a management consulting white paper.



- Describe at least one good feature of this visual display.
- Describe at least one bad feature of this visual display.
- Redraw the graph, using the guidelines given on page 74.

2.68 Professor Deanna Oxender Burgess of Florida Gulf Coast University conducted research on annual reports of corporations (see D. Rosato, "Worried About the Numbers? How About the Charts?" *The New York Times*, September 15, 2002, p. B7).

Burgess found that even slight distortions in a chart changed readers' perception of the information. Using online or library sources, select a corporation and study its most recent annual report. Find at least one chart in the report that you think needs improvement and develop an improved version of the chart. Explain why you believe the improved chart is better than the one included in the annual report.

2.69 Figure 2.1 shows a bar chart and a pie chart for the main reason young adults shop online (see page 51).

- Create an exploded pie chart, a doughnut chart, a cone chart, or a pyramid chart that shows the main reason young adults shop online.
- Which graphs do you prefer—the bar chart or pie chart or the exploded pie chart, doughnut chart, cone chart, and pyramid chart? Explain.

2.70 Figures 2.2 and 2.3 show a bar chart and a pie chart for the risk level for the retirement fund data (see page 52).

- Create an exploded pie chart, a doughnut chart, a cone chart, and a pyramid chart that shows the risk level of retirement funds.
- Which graphs do you prefer—the bar chart or pie chart or the exploded pie chart, doughnut chart, cone chart, and pyramid chart? Explain.

USING STATISTICS

The Choice Is Yours, Revisited

In the Using Statistics scenario, you were hired by the Choice Is Yours investment company to assist clients who seek to invest in retirement funds. A sample of 316 retirement funds was selected, and information on the funds and past performance history was recorded. For each of the 316 funds, data were collected on 13 variables. With so much information, visualizing all these numbers required the use of properly selected graphical displays.

From bar charts and pie charts, you were able to see that about two-thirds of the funds were classified as having low risk, about 30% had average risk, and about 4% had high risk. Contingency tables of the fund type and risk revealed that more of the value funds have low risk as compared to average or high. After constructing histograms and percentage polygons of the one-year returns, you were able to conclude that the one-year return was slightly higher for the value funds than for the growth funds. The return for both the growth and value funds is

concentrated between 10 and 20, the return for the growth funds is more concentrated between 10 and 15, and the return for the value funds is more concentrated between 15 and 20.

From a multidimensional contingency table, you discovered more complex relationships; for example, for the growth funds, the pattern of risk differs depending on the market cap of the fund.

With these insights, you can inform your clients about how the different funds performed. Of course, the past performance of a fund does not guarantee its future performance. You might also want to analyze the differences in return in the past three years, in the past 5 years, and the past 10 years to see how the growth funds, the value funds, and the small, mid-cap, and large market cap funds performed.



Dmitriy Shironosov/Shutterstock

SUMMARY

Organizing and visualizing data are the third and fourth tasks of the DCOVA framework. How you accomplish these tasks varies by the type of variable, categorical or

numerical, as well as the number of variables you seek to organize and visualize at the same time. Table 2.20 on page 76 summarizes the appropriate methods to do these tasks.

Using the appropriate methods to organize and visualize your data allows you to reach preliminary conclusions about the data. In several different chapter examples, tables and charts helped you reach conclusions about the main reason that young adults shop online and about the cost of restaurant meals in a city and its suburbs; they also provided some insights about the sample of retirement funds in The Choice Is Yours scenario.

Using the appropriate methods to visualize your data may help you reach preliminary conclusions as well as cause

you to ask additional questions about your data that may lead to further analysis at a later time. If used improperly, methods to organize and visualize the variables can obscure data or create false impressions, as Section 2.7 discusses.

Methods to organize and visualize data help summarize data. For numerical variables, there are many additional ways to summarize data that involve computing sample statistics or population parameters. The most common examples of these, *numerical descriptive measures*, are the subject of Chapter 3.

TABLE 2.20

Organizing and Visualizing Data

Type of Variable	Methods
Categorical variables	
Organize	Summary table, contingency table (Section 2.1)
Visualize one variable	Bar chart, pie chart, Pareto chart (Section 2.3)
Visualize two variables	Side-by-side chart (Section 2.3)
Numerical variables	
Organize	Ordered array, frequency distribution, relative frequency distribution, percentage distribution, cumulative percentage distribution (Section 2.2)
Visualize one variable	Stem-and-leaf display, histogram, percentage polygon, cumulative percentage polygon (ogive) (Section 2.4)
Visualize two variables	Scatter plot, time-series plot (Section 2.5)
Many variables together	
Organize	Multidimensional tables (Section 2.6)

REFERENCES

1. Batemen, S., R. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. "Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts." April 10, 2010, www.hci.usask.ca/uploads/173-pap0297-bateman.pdf.
2. Gross, Bertram. *The Managing of Organizations: The Administrative Struggle, Vols. I & II*. New York: The Free Press of Glencoe, 1964.
3. Huff, D. *How to Lie with Statistics*. New York: Norton, 1954.
4. Microsoft Excel 2013. Redmond, WA: Microsoft Corporation, 2012.
5. Minitab Release 16. State College, PA: Minitab, 2010.
6. Tufte, E. R. *Beautiful Evidence*. Cheshire, CT: Graphics Press, 2006.
7. Tufte, E. R. *Envisioning Information*. Cheshire, CT: Graphics Press, 1990.
8. Tufte, E. R. *The Visual Display of Quantitative Information*, 2nd ed. Cheshire, CT: Graphics Press, 2002.
9. Tufte, E. R. *Visual Explanations*. Cheshire, CT: Graphics Press, 1997.
10. Wainer, H. *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. New York: Copernicus/Springer-Verlag, 1997.

KEY EQUATIONS

Determining the Class Interval Width

$$\text{Interval width} = \frac{\text{highest value} - \text{lowest value}}{\text{number of classes}} \quad (2.1)$$

Computing the Proportion or Relative Frequency

$$\text{Proportion} = \text{relative frequency} = \frac{\text{number of values in each class}}{\text{total number of values}} \quad (2.2)$$

KEY TERMS

bar chart 51
 bins 45
 cell 39
 chartjunk 72
 class boundaries 43
 class interval 43
 class interval width 43
 class midpoints 44
 classes 43
 contingency table 39
 cumulative percentage distribution 47
 cumulative percentage polygon (ogive) 61

frequency distribution 43
 histogram 59
 joint response 39
 lurking variables 69
 multidimensional contingency table 68
 ogive (cumulative percentage polygon) 61
 ordered array 42
 Pareto chart 53
 Pareto principle 53
 percentage distribution 45
 percentage polygon 60
 pie chart 52

PivotTable 68
 proportion 45
 relative frequency 45
 relative frequency distribution 45
 scatter plot 65
 side-by-side bar chart 55
 stacked 49
 stem-and-leaf display 58
 summary table 38
 time-series plot 66
 unstacked 49

CHECKING YOUR UNDERSTANDING

2.71 How do histograms and polygons differ in construction and use?

2.72 Why would you construct a summary table?

2.73 What are the advantages and disadvantages of using a bar chart, a pie chart, and a Pareto chart?

2.74 Compare and contrast the bar chart for categorical data with the histogram for numerical data.

2.75 What is the difference between a time-series plot and a scatter plot?

2.76 Why is it said that the main feature of a Pareto chart is its ability to separate the “vital few” from the “trivial many”?

2.77 What are the three different ways to break down the percentages in a contingency table?

2.78 How can a multidimensional table differ from a two-variable contingency table?

2.79 What type of insights can you gain from a contingency table that contains three variables that you cannot gain from a contingency table that contains two variables?

CHAPTER REVIEW PROBLEMS

2.80 The following summary table presents the breakdown of the price of a new college textbook:

Revenue Category	Percentage (%)
Publisher	64.8
Manufacturing costs	32.3
Marketing and promotion	15.4
Administrative costs and taxes	10.0
After-tax profit	7.1
Bookstore	22.4
Employee salaries and benefits	11.3
Operations	6.6
Pretax profit	4.5
Author	11.6
Freight	1.2

Source: Data extracted from T. Lewin, “When Books Break the Bank,” *The New York Times*, September 16, 2003, pp. B1, B4.

- a. Using the four categories of publisher, bookstore, author, and freight, construct a bar chart, a pie chart, and a Pareto chart.
- b. Using the four subcategories of publisher and three subcategories of bookstore, along with the author and freight categories, construct a Pareto chart.
- c. Based on the results of (a) and (b), what conclusions can you reach concerning who gets the revenue from the sales of new college textbooks? Do any of these results surprise you? Explain.

2.81 The following table represents the market share (in number of movies, gross in millions of dollars, and millions of tickets sold) of each type of movie in 2012:

Type	Number	Gross (\$millions)	Tickets (millions)
Based on fiction book/short story	80	3181.1	541.8
Based on comic/graphic novel	9	1,552.3	198.2
Based on factual book/article	17	328.1	41.9
Based on game	3	125.1	15.9
Based on real life events	168	371.5	47.4
Based on TV	9	261.4	33.4
Original screenplay	367	4,245.0	541.8
Remake	14	194.3	24.7
Based on folk tale/legend/fairytales	4	304.5	38.9
Spin-off	2	39.9	5.1

Source: Data extracted from www.the-numbers.com/market/Sources2012.summary.

- a. Construct a bar chart, a pie chart, and a Pareto chart for the number of movies, gross (in \$millions), and number of tickets sold (in millions).
- b. What conclusions can you reach about the market shares of the different types of movies in 2012?

2.82 A survey was conducted from 665 consumer magazines on the practices of their websites. The results are summarized in a copyediting table and a fact-checking table:

Copyediting as Compared to Print Content	Percentage (%)
As rigorous	41
Less rigorous	48
Not copyedited	11

Source: Data extracted from S. Clifford, "Columbia Survey Finds a Slack Editing Process of Magazine Web Sites," *The New York Times*, March 1, 2010, p. B6.

- a. For copyediting, construct a bar chart, a pie chart, and a Pareto chart.
- b. Which graphical method do you think is best for portraying these data?

Fact Checking as Compared to Print Content	Percentage (%)
Same	57
Less rigorous	27
Online not fact-checked	8
Neither online nor print is fact-checked	8

Source: Data extracted from S. Clifford, "Columbia Survey Finds a Slack Editing Process of Magazine Web Sites," *The New York Times*, March 1, 2010, p. B6.

- c. For fact checking, construct a bar chart, a pie chart, and a Pareto chart.
- d. Which graphical method do you think is best for portraying the fact checking data?
- e. What conclusions can you reach concerning copyediting and fact checking of print and online consumer magazines?

2.83 The owner of a restaurant that serves Continental-style entrées has the business objective of learning more about the patterns of patron demand during the Friday-to-Sunday weekend time period. Data were collected from 630 customers on the type of entrée ordered and organized in the following table (and stored in **Entree**):

Type of Entrée	Number Served
Beef	187
Chicken	103
Mixed	30
Duck	25
Fish	122
Pasta	63
Shellfish	74
Veal	26
Total	630

- a. Construct a percentage summary table for the types of entrées ordered.
- b. Construct a bar chart, a pie chart, and a Pareto chart for the types of entrées ordered.
- c. Do you prefer using a Pareto chart or a pie chart for these data? Why?
- d. What conclusions can the restaurant owner reach concerning demand for different types of entrées?

2.84 Suppose that the owner of the restaurant in Problem 2.83 also wants to study the demand for dessert during the same time period. She decides that in addition to studying whether a dessert was ordered, she will also study the gender of the individual and whether a beef entrée was ordered. Data were collected from 630 customers and organized in the following contingency tables:

		GENDER	
		Male	Female
DESSERT ORDERED	Yes	50	96
	No	250	234
	Total	300	330
			630

		BEEF ENTRÉE	
		Yes	No
DESSERT ORDERED	Yes	74	68
	No	123	365
	Total	197	433
			630

- a. For each of the two contingency tables, construct contingency tables of row percentages, column percentages, and total percentages.

- b.** Which type of percentage (row, column, or total) do you think is most informative for each gender? For beef entrée? Explain.
c. What conclusions concerning the pattern of dessert ordering can the restaurant owner reach?

2.85 The following data represent the pounds per capita of fresh food and packaged food consumed in the United States, Japan, and Russia in a recent year:

	COUNTRY		
FRESH FOOD	United States	Japan	Russia
Eggs, nuts, and beans	88	94	88
Fruit	124	126	88
Meat and seafood	197	146	125
Vegetables	194	278	335
PACKAGED FOOD			
Bakery goods	108	53	144
Dairy products	298	147	127
Pasta	12	32	16
Processed, frozen, dried and chilled food, and ready-to-eat meals	183	251	70
Sauces, dressings, and condiments	63	75	49
Snacks and candy	47	19	24
Soup and canned food	77	17	25

Source: Data extracted from H. Fairfield, "Factory Food," *The New York Times*, April 4, 2010, p. BU5.

- a.** For the United States, Japan, and Russia, construct a bar chart, a pie chart, and a Pareto chart for different types of fresh foods consumed.
b. For the United States, Japan, and Russia, construct a bar chart, a pie chart, and a Pareto chart for different types of packaged foods consumed.
c. What conclusions can you reach concerning differences between the United States, Japan, and Russia in the fresh foods and packaged foods consumed?

2.86 The Air Travel Consumer Report, a monthly product of the Department of Transportation's Office of Aviation Enforcement and Proceedings (OAEP), is designed to assist consumers with information on the quality of services provided by airlines. The report includes a summary of consumer complaints by industry group and by complaint category. A breakdown of 987 November 2012 consumer complaints based on industry group is given in the following table:

Industry Group	Number of Consumer Complaints
Airlines	922
Travel agents	23
Tour operators	24
Miscellaneous	18
Industry total	987

Source: Data extracted from "The Travel Consumer Report," Office of Aviation Enforcement and Proceedings, January 2013.

- a.** Construct a Pareto chart for the number of complaints by industry group. What industry group accounts for most of the complaints?

The 922 consumer complaints against airlines fall into one of two groups: complaints against U.S. airlines and complaints against foreign airlines. The following table summarizes these 922 complaints by complaint type:

Complaint Type	Against U.S. Airlines	Against Foreign Airlines
Flight problems	201	38
Oversales	7	5
Reservation/ticketing/boarding	103	39
Fares	48	31
Refunds	47	13
Baggage	98	51
Customer service	100	38
Disability	48	6
Advertising	4	2
Discrimination	3	4
Animals	0	0
Other	29	7
Total	688	234

- b.** Construct pie charts to display the percentage of complaints by type against U.S. airlines and foreign airlines.
c. Construct a Pareto chart for the complaint categories against U.S. airlines. Does a certain complaint category account for most of the complaints?
d. Construct a Pareto chart for the complaint categories against foreign airlines. Does a certain complaint category account for most of the complaints?

2.87 One of the major measures of the quality of service provided by an organization is the speed with which the organization responds to customer complaints. A large family-held department store selling furniture and flooring, including carpet, had undergone a major expansion in the past several years. In particular, the flooring department had expanded from 2 installation crews to an installation supervisor, a measurer, and 15 installation crews. A business objective of the company was to reduce the time between when the complaint is received and when it is resolved. During a recent year, the company received 50 complaints concerning carpet installation. The number of days between the receipt of the complaint and the resolution of the complaint for the 50 complaints, stored in **Furniture**, are:

54	5	35	137	31	27	152	2	123	81	74	27
11	19	126	110	110	29	61	35	94	31	26	5
12	4	165	32	29	28	29	26	25	1	14	13
13	10	5	27	4	52	30	22	36	26	20	23
33	68										

- a.** Construct a frequency distribution and a percentage distribution.
b. Construct a histogram and a percentage polygon.
c. Construct a cumulative percentage distribution and plot a cumulative percentage polygon (ogive).

- d. On the basis of the results of (a) through (c), if you had to tell the president of the company how long a customer should expect to wait to have a complaint resolved, what would you say? Explain.

2.88 The file **DomesticBeer** contains the percentage alcohol, number of calories per 12 ounces, and number of carbohydrates (in grams) per 12 ounces for 152 of the best-selling domestic beers in the United States.

Source: Data extracted from www.beer100.com/beercalories.htm, March 20, 2013.

- Construct a percentage histogram for percentage alcohol, number of calories per 12 ounces, and number of carbohydrates (in grams) per 12 ounces.
- Construct three scatter plots: percentage alcohol versus calories, percentage alcohol versus carbohydrates, and calories versus carbohydrates.
- Discuss what you learn from studying the graphs in (a) and (b).

2.89 The file **CigaretteTax** contains the state cigarette tax (\$) for each state as of January 1, 2013.

- Construct an ordered array.
- Plot a percentage histogram.
- What conclusions can you reach about the differences in the state cigarette tax between the states?

2.90 The file **CDRate** contains the yields for one-year certificates of deposit (CDs) and a five-year CDs for 23 banks in the United States, as of March 20, 2013.

Source: Data extracted and compiled from www.Bankrate.com, March 20, 2013.

- Construct a stem-and-leaf display for one-year CDs and five-year CDs.
- Construct a scatter plot of one-year CDs versus five-year CDs.
- What is the relationship between the one-year CD rate and the five-year CD rate?

2.91 The file **CEO-Compensation** includes the total compensation (in \$millions) for CEOs of 170 large public companies and the investment return in 2012. (Data extracted from “CEO Pay Skyrockets as Economy, Stocks Recover,” *USA Today*, March 27, 2013, p. B1.) For total compensation:

- Construct a frequency distribution and a percentage distribution.
- Construct a histogram and a percentage polygon.
- Construct a cumulative percentage distribution and plot a cumulative percentage polygon (ogive).
- Based on (a) through (c), what conclusions can you reach concerning CEO compensation in 2012?
- Construct a scatter plot of total compensation and investment return in 2012.
- What is the relationship between the total compensation and investment return in 2012?

2.92 Studies conducted by a manufacturer of Boston and Vermont asphalt shingles have shown product weight to be a major factor in customers’ perception of quality. Moreover, the weight represents the amount of raw materials being used and is therefore very important to the company from a cost standpoint.

The last stage of the assembly line packages the shingles before the packages are placed on wooden pallets. The variable of interest is the weight in pounds of the pallet, which for most brands holds 16 squares of shingles. The company expects pallets of its Boston brand-name shingles to weigh at least 3,050 pounds but less than 3,260 pounds. For the company’s Vermont brand-name shingles, pallets should weigh at least 3,600 pounds but less than 3,800. Data, collected from a sample of 368 pallets of Boston shingles and 330 pallets of Vermont shingles, are stored in **Pallet**.

- For the Boston shingles, construct a frequency distribution and a percentage distribution having eight class intervals, using 3,015, 3,050, 3,085, 3,120, 3,155, 3,190, 3,225, 3,260, and 3,295 as the class boundaries.
- For the Vermont shingles, construct a frequency distribution and a percentage distribution having seven class intervals, using 3,550, 3,600, 3,650, 3,700, 3,750, 3,800, 3,850, and 3,900 as the class boundaries.
- Construct percentage histograms for the Boston shingles and for the Vermont shingles.
- Comment on the distribution of pallet weights for the Boston and Vermont shingles. Be sure to identify the percentages of pallets that are underweight and overweight.

2.93 What was the average price of a room at two-star, three-star, and four-star hotels in cities around the world in 2012? The file **HotelPrices** contains the prices in English pounds (about US\$1.56 as of July 2012). (Data extracted from bit.ly/Q0qxe4.) Complete the following for two-star, three-star, and four-star hotels:

- Construct a frequency distribution and a percentage distribution.
- Construct a histogram and a percentage polygon.
- Construct a cumulative percentage distribution and plot a cumulative percentage polygon (ogive).
- What conclusions can you reach about the cost of two-star, three-star, and four-star hotels?
- Construct separate scatter plots of the cost of two-star hotels versus three-star hotels, two-star hotels versus four-star hotels, and three-star hotels versus four-star hotels.
- What conclusions can you reach about the relationship of the price of two-star, three-star, and four-star hotels?

2.94 The file **Protein** contains calorie and cholesterol information for popular protein foods (fresh red meats, poultry, and fish).

Source: U.S. Department of Agriculture.

- Construct a percentage histogram for the number of calories.
- Construct a percentage histogram for the amount of cholesterol.
- What conclusions can you reach from your analyses in (a) and (b)?

2.95 The file **Natural Gas** contains the monthly average wellhead and residential price for natural gas (dollars per thousand cubic feet) in the United States from January 1, 2008, to January 1, 2013. (Data extracted from “U.S. Natural Gas Prices,” 1.usa.gov/qHDWNz, March 1, 2013.) For the wellhead price and the residential price:

- Construct a time-series plot.
- What pattern, if any, is present in the data?

- c. Construct a scatter plot of the wellhead price and the residential price.
- d. What conclusion can you reach about the relationship between the wellhead price and the residential price?

2.96 The following data (stored in **Drink**) represent the amount of soft drink in a sample of 50 consecutively filled 2-liter bottles. The results are listed horizontally in the order of being filled:

2.109 2.086 2.066 2.075 2.065 2.057 2.052 2.044 2.036 2.038
 2.031 2.029 2.025 2.029 2.023 2.020 2.015 2.014 2.013 2.014
 2.012 2.012 2.012 2.010 2.005 2.003 1.999 1.996 1.997 1.992
 1.994 1.986 1.984 1.981 1.973 1.975 1.971 1.969 1.966 1.967
 1.963 1.957 1.951 1.951 1.947 1.941 1.941 1.938 1.908 1.894

- a. Construct a time-series plot for the amount of soft drink on the *Y* axis and the bottle number (going consecutively from 1 to 50) on the *X* axis.
- b. What pattern, if any, is present in these data?
- c. If you had to make a prediction about the amount of soft drink filled in the next bottle, what would you predict?
- d. Based on the results of (a) through (c), explain why it is important to construct a time-series plot and not just a histogram, as was done in Problem 2.47 on page 64.

2.97 The file **Currency** contains the exchange rates of the Canadian dollar, the Japanese yen, and the English pound from 1980 to 2012, where the Canadian dollar, the Japanese yen, and the English pound are expressed in units per U.S. dollar.

- a. Construct time-series plots for the yearly closing values of the Canadian dollar, the Japanese yen, and the English pound.
- b. Explain any patterns present in the plots.
- c. Write a short summary of your findings.
- d. Construct separate scatter plots of the value of the Canadian dollar versus the Japanese yen, the Canadian dollar versus the English pound, and the Japanese yen versus the English pound.
- e. What conclusions can you reach concerning the value of the Canadian dollar, Japanese yen, and English pound in terms of the U.S. dollar?

2.98 A/B testing is a method used by businesses to test different designs and formats of a webpage to determine if a new web page is more effective than a current web page. Web designers tested a new call to action button on its web page. Every visitor to the web page was randomly shown either the original call-to-action button (the control) or the new variation. The metric used to measure success was the download rate: the number of people who downloaded the file divided by the number of people who saw that particular call-to-action button. Results of the experiment yielded the following:

Variations	Downloads	Visitors
Original call to action button	351	3,642
New call to action button	485	3,556

- a. Compute the percentage of downloads for the original call-to-action button and the new call-to-action button.

- b. Construct a bar chart of the percentage of downloads for the original call-to-action button and the new call-to-action button.
- c. What conclusions can you reach concerning the original call-to-action button and the new call-to-action button?

Web designers tested a new web design on its web page. Every visitor to the web page was randomly shown either the original web design (the control) or the new variation. The metric used to measure success was the download rate: the number of people who downloaded the file divided by the number of people who saw that particular web design. Results of the experiment yielded the following:

Variations	Downloads	Visitors
Original web design	305	3,427
New web design	353	3,751

- d. Compute the percentage of downloads for the original web design and the new web design.
- e. Construct a bar chart of the percentage of downloads for the original web design and the new web design.
- f. What conclusions can you reach concerning the original web design and the new web design?
- g. Compare your conclusions in (f) with those in (c).

Web designers now tested two factors simultaneously—the call-to-action button and the new web design. Every visitor to the web page was randomly shown one of the following:

Old call to action button with original web design
 New call to action button with original web design
 Old call to action button with new web design
 New call to action button with new web design

Again, the metric used to measure success was the download rate: the number of people who downloaded the file divided by the number of people who saw that particular call-to-action button and web design. Results of the experiment yielded the following:

Call to Action Button	Web Design	Downloaded	Declined	Total
Original	Original	83	917	1,000
New	Original	137	863	1,000
Original	New	95	905	1,000
New	New	170	830	1,000
Total		485	3,515	4,000

- h. Compute the percentage of downloads for each combination of call-to-action button and web design.
- i. What conclusions can you reach concerning the original call to action button and the new call to action button and the original web design and the new web design?
- j. Compare your conclusions in (i) with those in (c) and (g).

2.99 (Class Project) Have each student in the class respond to the question “Which carbonated soft drink do you most prefer?” so that the instructor can tally the results into a summary table.

- Convert the data to percentages and construct a Pareto chart.
- Analyze the findings.

2.100 (Class Project) Cross-classify each student in the class by gender (male, female) and current employment status (yes, no), so that the instructor can tally the results.

- Construct a table with either row or column percentages, depending on which you think is more informative.

- What would you conclude from this study?
- What other variables would you want to know regarding employment in order to enhance your findings?

REPORT WRITING EXERCISES

2.101 Referring to the results from Problem 2.92 on page 80 concerning the weights of Boston and Vermont shingles, write a report that evaluates whether the weights of the pallets of the two types of shingles are what the company expects. Be sure to incorporate tables and charts into the report.

CASES FOR CHAPTER 2

Managing Ashland MultiComm Services

Recently, Ashland MultiComm Services has been criticized for its inadequate customer service in responding to questions and problems about its telephone, cable television, and Internet services. Senior management has established a task force charged with the business objective of improving customer service. In response to this charge, the task force collected data about the types of customer service errors, the cost of customer service errors, and the cost of wrong billing errors. It found the following data:

Types of Customer Service Errors	
Type of Errors	Frequency
Incorrect accessory	27
Incorrect address	42
Incorrect contact phone	31
Invalid wiring	9
On-demand programming error	14
Subscription not ordered	8
Suspension error	15
Termination error	22
Website access error	30
Wrong billing	137
Wrong end date	17
Wrong number of connections	19
Wrong price quoted	20
Wrong start date	24
Wrong subscription type	33
Total	448

Cost of Customer Service Errors in the Past Year	
Type of Errors	Cost (\$ thousands)
Incorrect accessory	17.3
Incorrect address	62.4
Incorrect contact phone	21.3
Invalid wiring	40.8
On-demand programming errors	38.8
Subscription not ordered	20.3
Suspension error	46.8
Termination error	50.9
Website access errors	60.7
Wrong billing	121.7
Wrong end date	40.9
Wrong number of connections	28.1
Wrong price quoted	50.3
Wrong start date	40.8
Wrong subscription type	60.1
Total	701.2

Type and Cost of Wrong Billing Errors	
Type of Wrong Billing Errors	Cost (\$ thousands)
Declined or held transactions	7.6
Incorrect account number	104.3
Invalid verification	9.8
Total	121.7

1. Review these data (stored in **AMS2-1**). Identify the variables that are important in describing the customer service problems. For each variable you identify, construct the graphical representation you think is most appropriate and explain your choice. Also, suggest what other information concerning the different types of errors would be useful to examine. Offer possible courses of action for either the task force or management to take

that would support the goal of improving customer service.

2. As a follow-up activity, the task force decides to collect data to study the pattern of calls to the help desk (stored in **AMS2-2**). Analyze these data and present your conclusions in a report.

Digital Case

In the Using Statistics scenario, you were asked to gather information to help make wise investment choices. Sources for such information include brokerage firms, investment counselors, and other financial services firms. Apply your knowledge about the proper use of tables and charts in this Digital Case about the claims of foresight and excellence by an Ashland-area financial services firm.

Open **EndRunGuide.pdf**, which contains the EndRun Financial Services “Guide to Investing.” Review the guide, paying close attention to the company’s investment claims and supporting data and then answer the following.

1. How does the presentation of the general information about EndRun in this guide affect your perception of the business?

2. Is EndRun’s claim about having more winners than losers a fair and accurate reflection of the quality of its investment service? If you do not think that the claim is a fair and accurate one, provide an alternate presentation that you think is fair and accurate.
3. Review the discussion about EndRun’s “Big Eight Difference” and then open and examine the attached sample of mutual funds. Are there any other relevant data from that file that could have been included in the Big Eight table? How would the new data alter your perception of EndRun’s claims?
4. EndRun is proud that all Big Eight funds have gained in value over the past five years. Do you agree that EndRun should be proud of its selections? Why or why not?

CardioGood Fitness

The market research team at AdRight is assigned the task to identify the profile of the typical customer for each treadmill product offered by CardioGood Fitness. The market research team decides to investigate whether there are differences across the product lines with respect to customer characteristics. The team decides to collect data on individuals who purchased a treadmill at a CardioGood Fitness retail store during the prior three months. The data are stored in the **CardioGood Fitness** file. The team identifies the following customer variables to study: product purchased, TM195, TM498, or TM798; gender; age, in years; education, in years; relationship status, single or partnered; annual

household income (\$); average number of times the customer plans to use the treadmill each week; average number of miles the customer expects to walk/run each week; and self-rated fitness on an 1-to-5 ordinal scale, where 1 is poor shape and 5 is excellent shape.

1. Create a customer profile for each CardioGood Fitness treadmill product line by developing appropriate tables and charts.
2. Write a report to be presented to the management of CardioGood Fitness detailing your findings.

The Choice Is Yours Follow-Up

Follow up the Using Statistics Revisited section on page 75 by analyzing the differences in 3-year return percentages, 5-year return percentages, and 10-year return percentages for the sample of 316 retirement funds stored in

Retirement Funds. In your analysis, examine differences between the growth and value funds as well as the differences among the small, mid-cap, and large market cap funds.

Clear Mountain State Student Surveys

1. The student news service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. They create and distribute a survey of 14 questions and receive responses from 62 undergraduates (stored in **UndergradSurvey**). For each question asked in the survey, construct all the appropriate tables and charts and write a report summarizing your conclusions.
2. The dean of students at CMSU has learned about the undergraduate survey and has decided to undertake a similar survey for graduate students at CMSU. She creates and distributes a survey of 14 questions and receives responses from 44 graduate students (stored in **GradSurvey**). For each question asked in the survey, construct all the appropriate tables and charts and write a report summarizing your conclusions.

CHAPTER 2 EXCEL GUIDE

EG2.1 ORGANIZING CATEGORICAL VARIABLES

The Summary Table

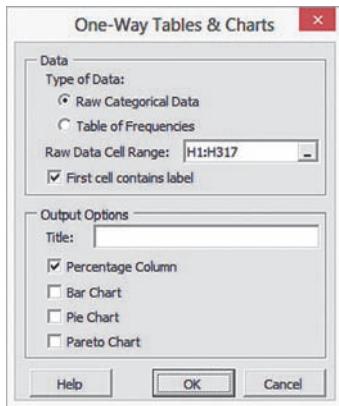
Key Technique Use the PivotTable feature to create a summary table for untallied data.

Example Create a frequency and percentage summary table similar to Table 2.3 on page 38.

PHStat Use One-Way Tables & Charts.

For the example, open to the **DATA worksheet** of the **Retirement Funds workbook**. Select **PHStat → Descriptive Statistics → One-Way Tables & Charts**. In the procedure's dialog box (shown below):

1. Click **Raw Categorical Data** (because the worksheet contains untallied data).
2. Enter H1:H317 as the **Raw Data Cell Range** and check **First cell contains label**.
3. Enter a **Title**, check **Percentage Column**, and click **OK**.



PHStat creates a PivotTable summary table on a new worksheet. For data that have already been tallied into categories, click **Table of Frequencies** in step 1.

In the PivotTable, risk categories appear in alphabetical order and not in the order low, average, and high as would normally be expected. To change to the expected order, use steps 14 and 15 of the *In-Depth Excel* instructions but change all references to cell A6 to cell A7 and drop the Low label over cell A5, not cell A4.

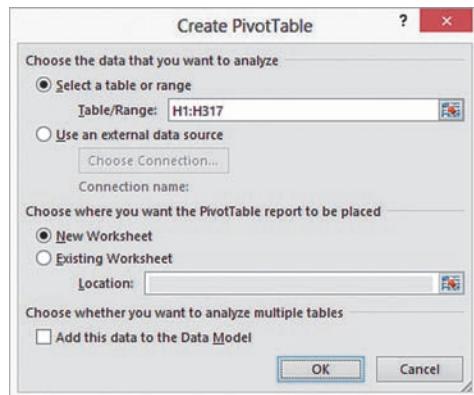
In-Depth Excel (untallied data)

Use the **Summary Table workbook** as a model.

For the example, open to the **DATA worksheet** of the **Retirement Funds workbook** and select **Insert → PivotTable**. In the Create PivotTable dialog box (shown at top in right column):

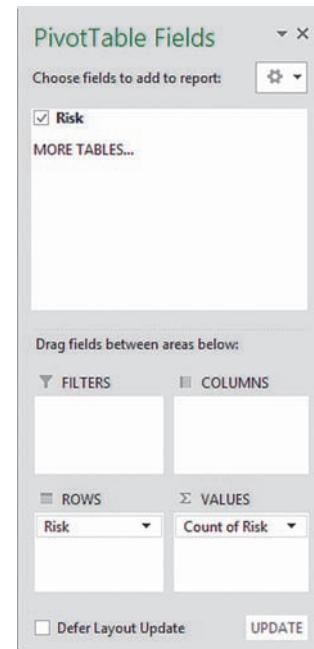
1. Click **Select a table or range** and enter H1:H317 as the Table/Range cell range.

2. Click **New Worksheet** and then click **OK**.



In the Excel 2013 PivotTable Fields task pane (shown below) or in the similar PivotTable Field List task pane in other Excels:

3. Drag **Risk** in the **Choose fields to add to report box** and drop it in the **ROWS (or Row Labels)** box.
4. Drag **Risk** in the **Choose fields to add to report box** a second time and drop it in the **Σ Values** box. This second label changes to **Count of Risk** to indicate that a count, or tally, of the risk categories will be displayed in the PivotTable.

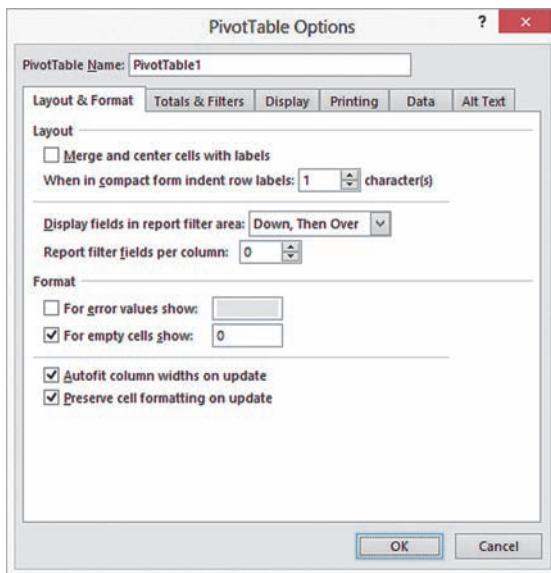


In the PivotTable being created:

5. Enter **Risk** in cell **A3** to replace the heading Row Labels.
6. Right-click cell **A3** and then click **PivotTable Options** in the shortcut menu that appears.

In the PivotTable Options dialog box (shown below):

7. Click the **Layout & Format** tab.
8. Check **For empty cells show** and enter **0** as its value. Leave all other settings unchanged.
9. Click **OK** to complete the PivotTable.



To add a column for the percentage frequency:

10. Enter **Percentage** in cell **C3**. Enter the formula $=B4/B\$7$ in cell **C4** and copy it down through **row 7**.
11. Select cell range **C4:C7**, right-click, and select **Format Cells** in the shortcut menu.
12. In the **Number** tab of the Format Cells dialog box, select **Percentage** as the **Category** and click **OK**.
13. Adjust the worksheet formatting, if appropriate (see Appendix B) and enter a title in cell **A1**.

In the PivotTable, risk categories appear in alphabetical order and not in the order low, average, and high, as would normally be expected. To change to the expected order:

14. Click the **Low** label in cell **A6** to highlight cell A6. Move the mouse pointer to the top edge of the cell until the mouse pointer changes to a four-way arrow.
15. Drag the **Low** label and drop the label over cell **A4**. The risk categories now appear in the order Low, Average, and High in the summary table.

In-Depth Excel (tallied data) Use the **SUMMARY_SIMPLE** worksheet of the **Summary Table workbook** as a model for creating a summary table.

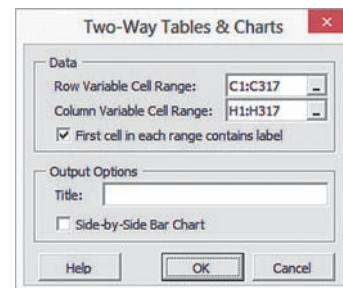
The Contingency Table

Key Technique Use the PivotTable feature to create a contingency table for untallied data.

Example Construct a contingency table displaying fund type and risk level similar to Table 2.4 on page 39.

PHStat (untallied data) Use **Two-Way Tables & Charts**. For the example, open to the **DATA worksheet** of the **Retirement Funds workbook**. Select **PHStat → Descriptive Statistics → Two-Way Tables & Charts**. In the procedure's dialog box (shown below):

1. Enter **C1:C317** as the **Row Variable Cell Range**.
2. Enter **H1:H317** as the **Column Variable Cell Range**.
3. Check **First cell in each range contains label**.
4. Enter a **Title** and click **OK**.



In the PivotTable, risk categories appear in alphabetical order and not in the order low, average, and high as would normally be expected. To change the expected order, use steps 14 and 15 of the *In-Depth Excel* instructions.

In-Depth Excel (untallied data) Use the **Contingency Table workbook** as a model.

For the example, open to the **DATA worksheet** of the **Retirement Funds workbook**. Select **Insert → PivotTable**. In the Create PivotTable dialog box:

1. Click **Select a table or range** and enter **A1:N317** as the **Table/Range** cell range.

2. Click **New Worksheet** and then click **OK**.

In the PivotTable Fields (called the PivotTable Field List in some Excel versions) task pane:

3. Drag **Type** from **Choose fields to add to report** and drop it in the **ROWS** (or **Row Labels**) box.
4. Drag **Risk** from **Choose fields to add to report** and drop it in the **COLUMNS** (or **Column Labels**) box.
5. Drag **Type** from **Choose fields to add to report** a second time and drop it in the **Σ VALUES** box. (**Type** changes to **Count of Type**.)

In the PivotTable being created:

6. Select cell **A3** and enter a **space character** to clear the label **Count of Type**.
7. Enter **Type** in cell **A4** to replace the heading Row Labels.
8. Enter **Risk** in cell **B3** to replace the heading Column Labels.
9. Click the **Low** label in cell **D4** to highlight cell D4. Move the mouse pointer to the left edge of the cell until the mouse pointer changes to a four-way arrow.
10. Drag the **Low** label to the left and drop the label when an I-beam appears between columns A and B. The **Low** label appears in **B4** and column B now contains the low risk tallies.

11. Right-click over the PivotTable and then click **PivotTable Options** in the shortcut menu that appears.

In the PivotTable Options dialog box:

12. Click the **Layout & Format** tab.
13. Check **For empty cells show** and enter **0** as its value. Leave all other settings unchanged.
14. Click the **Total & Filters** tab.
15. Check **Show grand totals for columns** and **Show grand totals for rows**.
16. Click **OK** to complete the table.

In-Depth Excel (tallied data) Use the **CONTINGENCY_SIMPLE** worksheet of the **Contingency Table** workbook as a model for creating a contingency table.

EG2.2 ORGANIZING NUMERICAL VARIABLES

Stacked and Unstacked Data

PHStat Use Stack Data or Unstack Data.

For example, to unstack the **3YrReturn%** variable by the **Type** variable in the retirement funds sample, open to the **DATA worksheet** of the **Retirement Funds** workbook. Select **Data Preparation** → **Unstack Data**. In that procedure's dialog box, enter **C1:C317** (the **Type** variable cell range) as the **Grouping Variable Cell Range** and enter **J1:J317** (the **3YrReturn%** variable cell range) as the **Stacked Data Cell Range**. Check **First cells in both ranges contain label** and click **OK**. The unstacked data appear on a new worksheet.

The Ordered Array

In-Depth Excel To create an ordered array, first select the numerical variable to be sorted. Then select **Home** → **Sort & Filter** (in the Editing group) and in the drop-down menu click **Sort Smallest to Largest**. (You will see **Sort A to Z** as the first drop-down choice if you did not select a cell range of *numerical* data.)

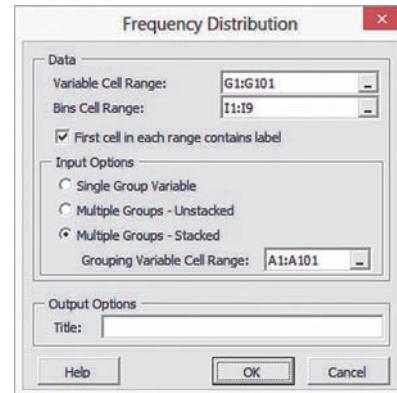
The Frequency Distribution

Key Technique Establish bins (see *Classes and Excel Bins* in Section 2.2) and then use the **FREQUENCY(untallied data cell range, bins cell range)** array function to tally data.

Example Create a frequency, percentage, and cumulative percentage distribution for the restaurant meal cost data that contains the information found in Tables 2.9, 2.11, and 2.14, in Section 2.2.

PHStat (untallied data) Use **Frequency Distribution**. (Use **Histogram & Polygons**, discussed in Section EG2.4, if you plan to construct a histogram or polygon in addition to a frequency distribution.) For the example, open to the **DATA worksheet** of the **Restaurants** workbook. This worksheet contains the meal cost data in stacked format in column G and a set of bin numbers appropriate for those data in column H. Select **PHStat** → **Descriptive Statistics** → **Frequency Distribution**. In the procedure's dialog box (shown below):

1. Enter **G1:G101** as the **Variable Cell Range**, enter **I1:I9** as the **Bins Cell Range**, and check **First cell in each range contains label**.
2. Click **Multiple Groups - Stacked** and enter **A1:A101** as the **Grouping Variable Cell Range**. (The cell range A1:A101 contains the Location variable.)
3. Enter a **Title** and click **OK**.



Click **Single Group Variable** in step 2 if constructing a distribution from a single group of untallied data. Click **Multiple Groups - Unstacked** in step 2 if the **Variable Cell Range** contains two or more columns of unstacked, untallied data.

Frequency distributions for the two groups appear on separate worksheets. To display the information for the two groups on one worksheet, select the cell range **B3:D11** on one of the worksheets. Right-click that range and click **Copy** in the shortcut menu. Open to the other worksheet. In that other worksheet, right-click cell **E3** and click **Paste Special** in the shortcut menu. In the Paste Special dialog box, click **Values and numbers format** and click **OK**. Adjust the worksheet title as necessary. (Learn more about the Paste Special command in Appendix B.)

In-Depth Excel (untallied data) Use the **Distributions** workbook as a model.

For the example, open to the **UNSTACKED** worksheet of the **Restaurants** workbook. This worksheet contains the meal cost data unstacked in columns A and B and a set of bin numbers appropriate for those data in column D. Then:

1. Right-click the **UNSTACKED** sheet tab and click **Insert** in the shortcut menu.
2. In the **General** tab of the Insert dialog box, click **Worksheet** and then click **OK**.

In the new worksheet:

3. Enter a title in cell **A1**, **Bins** in cell **A3**, and **Frequency** in cell **B3**.
4. Copy the bin number list in the cell range **D2:D9** of the **UNSTACKED** worksheet and paste this list into cell **A4** of the new worksheet.
5. Select the cell range **B4:B12** that will hold the array formula.

6. Type, but do not press the **Enter** or **Tab** key, the formula **=FREQUENCY(UNSTACKED!\$A\$1:\$A\$51, \$A\$4:\$A\$11)**. Then, while holding down the **Ctrl** and **Shift** keys, press the **Enter** key to enter the array formula into the cell range **B4:B12**. (Learn more about array formulas in Appendix B.)

7. Adjust the worksheet formatting as necessary.

Note that in step 6, you enter the cell range as **UNSTACKED!\$A\$1:\$A\$51** and not as **\$A\$1:\$A\$51** because the untallied data are located on another (the UNSTACKED) worksheet. (Learn more about referring to data on another worksheet, as well as the significance of entering the cell range as **\$A\$1:\$A\$51** and not as **A1:A51**, in Appendix B.)

Steps 1 through 7 construct a frequency distribution for the meal costs at city restaurants. To construct a frequency distribution for the meal costs at suburban restaurants, repeat steps 1 through 7 but in step 6 type **=FREQUENCY(UNSTACKED!\$B\$1:\$B\$51, \$A\$4:\$A\$11)** as the array formula.

To display the distributions for the two groups on one worksheet, select the cell range **B3:B11** on one of the worksheets. Right-click that range and click **Copy** in the shortcut menu. Open to the other worksheet, right-click cell **C3** and click **Paste Special** in the shortcut menu. In the Paste Special dialog box, click **Values and numbers format** and click **OK**. Adjust the worksheet title as necessary. (Learn more about the Paste Special command in Appendix B.)

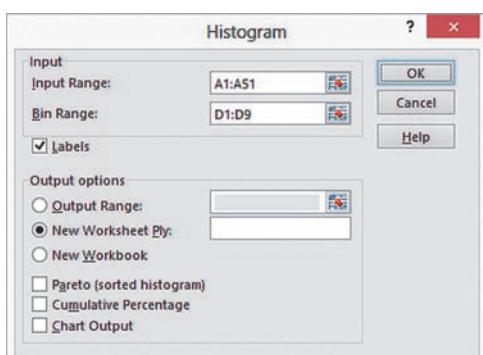
Analysis ToolPak (untallied data) Use Histogram.

For the example, open to the **UNSTACKED worksheet** of the **Restaurants workbook**. This worksheet contains the meal cost data unstacked in columns A and B and a set of bin numbers appropriate for those data in column D. Then:

1. Select **Data → Data Analysis**. In the Data Analysis dialog box, select **Histogram** from the **Analysis Tools** list and then click **OK**.

In the Histogram dialog box (shown below):

2. Enter **A1:A51** as the **Input Range** and enter **D1:D9** as the **Bin Range**. (If you leave **Bin Range** blank, the procedure creates a set of bins that will not be as well formed as the ones you can specify.)
3. Check **Labels** and click **New Worksheet Ply**.
4. Click **OK** to create the frequency distribution on a new worksheet.



In the new worksheet:

5. Select **row 1**. Right-click this row and click **Insert** in the shortcut menu. Repeat. (This creates two blank rows at the top of the worksheet.)

6. Enter a title in cell **A1**.

The ToolPak creates a frequency distribution that contains an improper bin labeled **More**. Correct this error by using these general instructions:

7. Manually add the frequency count of the **More** row to the frequency count of the preceding row. (For the example, the More row contains a zero for the frequency, so the frequency of the preceding row does not change.)
8. Select the worksheet row (for this example, row 12) that contains the **More** row.
9. Right-click that row and click **Delete** in the shortcut menu.

Steps 1 through 9 construct a frequency distribution for the meal costs at city restaurants. To construct a frequency distribution for the meal costs at suburban restaurants, repeat these nine steps but in step 6 enter **B1:B51** as the **Input Range**.

The Relative Frequency, Percentage, and Cumulative Distributions

Key Technique Add columns that contain formulas for the relative frequency or percentage and cumulative percentage to a previously constructed frequency distribution.

Example Create a distribution that includes the relative frequency or percentage as well as the cumulative percentage information found in Tables 2.11 (relative frequency and percentage) and 2.14 (cumulative percentage) in Section 2.2 for the restaurant meal cost data.

PHStat (untallied data) Use Frequency Distribution.

For the example, use the **PHStat** instructions in “The Frequency Distribution” to construct a frequency distribution. Note that the frequency distribution constructed by PHStat also includes columns for the percentages and cumulative percentages. To change the column of percentages to a column of relative frequencies, reformat that column. For the example, open to the new worksheet that contains the city restaurant frequency distribution and:

1. Select the cell range **C4:C11**, right-click, and select **Format Cells** from the shortcut menu.
2. In the **Number** tab of the Format Cells dialog box, select **Number** as the **Category** and click **OK**.

Then repeat these two steps for the new worksheet that contains the suburban restaurant frequency distribution.

In-Depth Excel (untallied data) Use the Distributions workbook as a model.

For the example, first construct a frequency distribution created using the **In-Depth Excel** instructions in “The Frequency

Distribution.” Open to the new worksheet that contains the frequency distribution for the city restaurants and:

1. Enter **Percentage** in cell **C3** and **Cumulative Pctage** in cell **D3**.
2. Enter $=B4/SUM($B$4:$B$11)$ in cell **C4** and copy this formula down through **row 11**.
3. Enter $=C4$ in cell **D4**.
4. Enter $=C5 + D4$ in cell **D5** and copy this formula down through row **11**.
5. Select the cell range **C4:D11**, right-click, and click **Format Cells** in the shortcut menu.
6. In the **Number** tab of the Format Cells dialog box, click **Percentage** in the **Category** list and click **OK**.

Then open to the worksheet that contains the frequency distribution for the suburban restaurants and repeat steps 1 through 6.

If you want column C to display relative frequencies instead of percentages, enter **Rel. Frequencies** in cell **C3**. Select the cell range **C4:C12**, right-click, and click **Format Cells** in the shortcut menu. In the **Number** tab of the Format Cells dialog box, click **Number** in the **Category** list and click **OK**.

Analysis ToolPak Use **Histogram** and then modify the worksheet created.

For the example, first construct the frequency distributions using the *Analysis ToolPak* instructions in “The Frequency Distribution.” Then use the *In-Depth Excel* instructions to modify those distributions.

EG2.3 VISUALIZING CATEGORICAL VARIABLES

Many of the *In-Depth Excel* instructions in the rest of this Excel Guide refer to the following labeled Charts group illustration.



The Bar Chart and the Pie Chart

Key Technique Use the Excel bar or pie chart feature. If the variable to be visualized is untallied, first construct a summary table (see the Section EG2.1 “The Summary Table” instructions).

Example Construct a bar or pie chart from a summary table similar to Table 2.3 on page 38.

PHStat Use **One-Way Tables & Charts**.

For the example, use the Section EG2.1 “The Summary Table” **PHStat** instructions, but in step 3, check either **Bar Chart** or **Pie Chart** (or both) in addition to entering a **Title**, checking **Percentage Column**, and clicking **OK**.

In-Depth Excel Use the **Summary Table workbook** as a model.

For the example, open to the **OneWayTable worksheet** of the **Summary Table workbook**. (The PivotTable in this worksheet

was constructed using the Section EG2.1 “The Summary Table” instructions.) To construct a bar chart:

1. Select cell range **A4:B6**. (Begin your selection at cell B6 and not at cell A4, as you would normally do.)
2. In Excel 2013, select **Insert**, then the **Bar icon** in the **Charts group** (#2 in the Charts group illustration), and then select the **first 2-D Bar** gallery item (**Clustered Bar**). In other Excels, select **Insert → Bar** and then select the **first 2-D Bar** gallery item (**Clustered Bar**).
3. Right-click the **Risk** drop-down button in the chart and click **Hide All Field Buttons on Chart**.
4. (Excel 2013) Select **Design → Add Chart Element → Axis Titles → Primary Horizontal**. (Other Excels) Select **Layout → Axis Titles → Primary Horizontal Axis Title → Title Below Axis**. Select the words “Axis Title” in the chart and enter the title **Frequency**.
5. Relocate the chart to a chart sheet and turn off the chart legend and gridlines by using the instructions in Appendix Section B.6.

Although not the case with the example, sometimes the horizontal axis scale of a bar chart will not begin at 0. If this occurs, right-click the horizontal (value) axis in the bar chart and click **Format Axis** in the shortcut menu. In the Excel 2013 Format Axis task pane, click **Axis Options**. In the Axis Options, enter **0** in the **Minimum** box and then close the pane. In other Excels, you set this value in the Format Axis dialog box. Click **Axis Options** in the left pane, and in the Axis Options right pane, click the first **Fixed** option button (for Minimum), enter **0** in its box, and then click **Close**.

To construct a pie chart, replace steps 2, 4, and 5 with these steps:

2. In Excel 2013, select **Insert**, then the **Pie icon** (#4 in the Step 4 illustration), and then select the **first 2-D Pie** gallery item (**Pie**). In other Excels, select **Insert → Pie** and then select the **first 2-D Pie** gallery item (**Pie**).
4. (Excel 2013) Select **Design → Add Chart Element → Data Labels → More Data Label Options**. In the Format Data Labels task pane, click **Label Options**. In the Label Options, check **Category Name** and **Percentage**, clear the other Label Contains check boxes, and click **Outside End**. (To see the Label Options, you may have to first click the chart [fourth] icon near the top of the task pane.) Then close the task pane. (Other Excels) Select **Layout → Data Labels → More Data Label Options**. In the Format Data Labels dialog box, click **Label Options** in the left pane. In the Label Options right pane, check **Category Name** and **Percentage** and clear the other Label Contains check boxes. Click **Outside End** and then click **Close**.
5. Relocate the chart to a chart sheet and turn off the chart legend and gridlines by using the instructions in Appendix Section B.6.

The Pareto Chart

Key Technique Use the Excel chart feature with a modified summary table.

Example Construct a Pareto chart of the incomplete ATM transactions equivalent to Figure 2.4 on page 54.

PHStat Use One-Way Tables & Charts.

For the example, open to the **DATA worksheet** of the **ATM Transactions workbook**. Select **PHStat → Descriptive Statistics → One-Way Tables & Charts**. In the procedure's dialog box:

1. Click **Table of Frequencies** (because the worksheet contains tallied data).
2. Enter **A1:B8** as the **Freq. Table Cell Range** and check **First cell contains label**.
3. Enter a **Title**, check **Pareto Chart**, and click **OK**.

In-Depth Excel Use the **Pareto workbook** as a model.

For the example, open to the **ATMTable worksheet** of the **ATM Transactions workbook**. Begin by sorting the modified table by decreasing order of frequency:

1. Select row **11** (the Total row), right-click, and click **Hide** in the shortcut menu. (This prevents the total row from getting sorted.)
2. Select cell **B4** (the first frequency), right-click, and select **Sort → Sort Largest to Smallest**.
3. Select rows **10** and **12** (there is no row 11 visible), right-click, and click **Unhide** in the shortcut menu to restore row 11.

Next, add a column for cumulative percentage:

4. Enter **Cumulative Pct.** in cell **D3**. Enter **=C4** in cell **D4**. Enter **=D4 + C5** in cell **D5** and copy this formula down through **row 10**.
5. Adjust the formatting of column D as necessary.

Next, create the Pareto chart:

6. Select the cell range **A3:A10** and while holding down the **Ctrl** key also select the cell range **C3:D10**.
7. In Excel 2013, select **Insert**, then the **Column icon** (#1 in the illustration on page 89), and select the **first 2-D Column** gallery item (**Clustered Column**). In other Excels, select **Insert → Column** and select the **first 2-D Column** gallery item (**Clustered Column**).
8. Select **Format**. In the **Current Selection** group, select **Series “Cumulative Pct.”** from the drop-down list and then click **Format Selection**.
9. (Excel 2013) In the Format Data Series task pane, click **Series Options**. In the Series Options, click **Secondary Axis**, and then close the task pane. (To see the Series Options, you may have to first click the chart [third] icon near the top of the task pane.)

(Other Excels) In the Format Data Series dialog box, click **Series Options** in the left pane, and in the **Series Options** right pane, click **Secondary Axis**. Click **Close**.

10. With the cumulative percentage series still selected in the Current Selection group, select **Design → Change Chart Type**. In Excel 2013, in the Change Chart Type dialog box, click **Combo** in the **All Charts** tab. In the **Cumulative Pct.** drop-down list, select the **fourth Line** gallery item (**Line with Markers**). Then, check **Secondary Axis** for the Cumulative Pct. and click **OK**. In other Excels, in the Change Chart Type dialog box, select the **fourth Line** gallery item (**Line with Markers**) and click **OK**.

Next, set the maximum value of the primary and secondary (left and right) Y axis scales to 100%. For each Y axis:

11. Right-click on the axis and click **Format Axis** in the shortcut menu.
12. (Excel 2013) In the Format Axis task pane, click **Axis Options**. In the Axis Options, enter **1** in the **Maximum** box. Click **Tick Marks** and select **Outside** from the **Major** type drop-down list. Then close the Format Axis pane. (To see the Axis Options, you may have to first click the chart [fourth] icon near the top of the task pane.)
- (Other Excels) In the Format Axis dialog box, click **Axis Options** in the left pane, and in the **Axis Options** right pane, click the **Fixed** option button for Maximum, enter **1** in its box, and click **Close**.
13. Relocate the chart to a chart sheet, turn off the chart legend and gridlines, and add chart and axis titles by using the instructions in Appendix Section B.6.

If you use a PivotTable as a summary table, replace steps 1 through 6 with these steps:

1. Add a percentage column in column C. (See the Section EG2.1 “The Summary Table” instructions, Steps 10 through 13.)
2. Add a cumulative percentage column in column D. Enter **Cumulative Pctage** in cell **D3**. Enter **=C4** in cell **D4**. Enter **=C5 + D4** in cell **D5**, and copy the formula down through all the rows in the PivotTable.
3. Select the total row, right-click, and click **Hide** in the shortcut menu. (This prevents the total row from getting sorted.)
4. Right-click the cell that contains the first frequency (typically this will be cell **B4**).
5. Right-click and select **Sort → Sort Largest to Smallest**.
6. Select the cell range of only the percentage and cumulative percentage columns (the equivalent of the cell range C3:D10 in the example).

The Pareto chart constructed from a PivotTable using these modified steps will not have proper labels for the categories. To add the correct labels, right-click over the chart and click **Select Data** in the shortcut menu. In the Select Data Source dialog box, click **Edit** that appears under **Horizontal (Category) Axis Labels**. In the Axis Labels dialog box, drag the mouse to *select* the cell range (**A4:A10** in the example) to enter that cell range. Do *not* type the cell range in the **Axis label range** box as you would otherwise do for the reasons explained in Appendix Section B.7. Click **OK** in this dialog box and then click **OK** in the original dialog box.

The Side-by-Side Chart

Key Technique Use an Excel bar chart that is based on a contingency table.

Example Construct a side-by-side chart that displays the fund type and risk level, similar to Figure 2.6 on page 55.

PHStat Use Two-Way Tables & Charts.

For the example, use the Section EG2.1 “The Contingency Table” **PHStat** instructions, but in step 4, check **Side-by-Side Bar Chart** in addition to entering a **Title** and clicking **OK**.

In-Depth Excel Use the **Contingency Table workbook** as a model.

For the example, open to the **TwoWayTable worksheet** of the **Contingency Table workbook** and:

1. Select cell A3 (or any other cell inside the PivotTable).
2. Select **Insert → Bar** and select the **first 2-D Bar** gallery item (**Clustered Bar**).
3. Right-click the **Risk** drop-down button in the chart and click **Hide All Field Buttons on Chart**.
4. Relocate the chart to a chart sheet, turn off the gridlines, and add chart and axis titles by using the instructions in Appendix Section B.6.

When creating a chart from a contingency table that is not a PivotTable, select the cell range of the contingency table, including row and column headings, but excluding the total row and total column, as step 1.

If you need to switch the row and column variables in a side-by-side chart, right-click the chart and then click **Select Data** in the shortcut menu. In the Select Data Source dialog box, click **Switch Row/Column** and then click **OK**. (In Excel 2007, if the chart is based on a PivotTable, the **Switch Row/Column** as that button will be disabled. In that case, you need to change the PivotTable to change the chart.)

EG2.4 VISUALIZING NUMERICAL VARIABLES

The Stem-and-Leaf Display

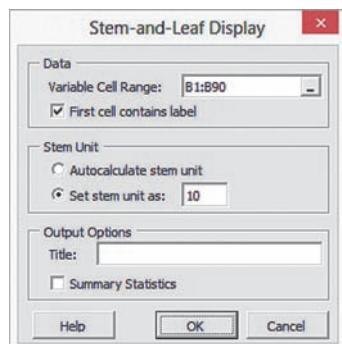
Key Technique Enter leaves as a string of digits that begin with the ' (apostrophe) character.

Example Construct a stem-and-leaf display of the one-year return percentage for the value retirement funds, similar to Figure 2.7 on page 58.

PHStat Use the **Stem-and-Leaf Display**.

For the example, open to the **UNSTACKED worksheet** of the **Retirement Funds workbook**. Select **PHStat → Descriptive Statistics → Stem-and-Leaf Display**. In the procedure's dialog box (shown in the next column):

1. Enter **B1:B90** as the **Variable Cell Range** and check **First cell contains label**.
2. Click **Set stem unit as** and enter **10** in its box.
3. Enter a **Title** and click **OK**.



When creating other displays, use the **Set stem unit as** option sparingly and only if **Autocalculate stem unit** creates a display that has too few or too many stems. (Any stem unit you specify must be a power of 10.)

In-Depth Excel Use the **Stem-and-leaf workbook** as a model. Manually construct the stems and leaves on a new worksheet to create a stem-and-leaf display. Adjust the column width of the column that holds the leaves as necessary.

The Histogram

Key Technique Modify an Excel column chart.

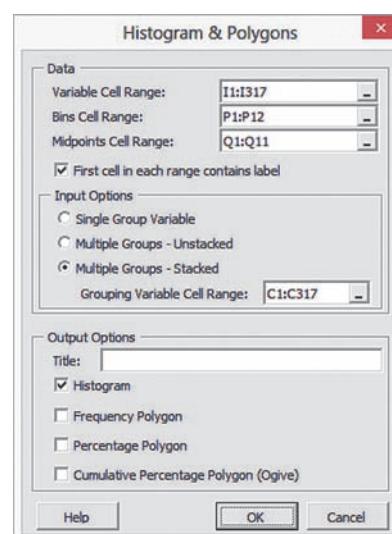
Example Construct histograms for the one-year return percentages for the growth and value retirement funds, similar to Figure 2.9 on page 60.

PHStat Use **Histogram & Polygons**.

For the example, open to the **DATA worksheet** of the **Retirement Funds workbook**. Select **PHStat → Descriptive Statistics → Histogram & Polygons**. In the procedure's dialog box (shown below):

1. Enter **I1:I317** as the **Variable Cell Range**, **P1:P12** as the **Bins Cell Range**, **Q1:Q11** as the **Midpoints Cell Range**, and check **First cell in each range contains label**.
2. Click **Multiple Groups - Stacked** and enter **C1:C317** as the **Grouping Variable Cell Range**. (In the DATA worksheet, the one-year return percentages for both types of retirement funds are stacked, or placed in a single column. The column C values allow PHStat to separate the returns for growth funds from the returns for the value funds.)
3. Enter a **Title**, check **Histogram**, and click **OK**.

PHStat inserts two new worksheets, each of which contains a frequency distribution and a histogram. To relocate the histograms to their own chart sheets, use the instructions in Appendix Section B.6.



As explained in Section 2.2, you cannot define an explicit lower boundary for the first bin, so the first bin can never have a midpoint. Therefore, the **Midpoints Cell Range** you enter must have one fewer cell than the **Bins Cell Range**. PHStat associates the first midpoint with the second bin and uses--as the label for the first bin.

The example uses the workaround discussed in “Classes and Excel Bins” in Section 2.2. When you use this workaround, the histogram bar labeled--will *always* be a zero bar. Appendix Section B.8 explains how you can delete this unnecessary bar from the histogram, as was done for the examples shown in Section 2.4.

In-Depth Excel

Use the **Histogram workbook** as a model. For the example, first construct frequency distributions for the growth and value funds. Open to the **UNSTACKED worksheet** of the **Retirement Funds workbook**. This worksheet contains the retirement funds data unstacked in columns A and B and a set of bin numbers and midpoints appropriate for those variables in columns D and E. Then:

1. Right-click the **UNSTACKED** sheet tab and click **Insert** in the shortcut menu.
2. In the **General** tab of the Insert dialog box, click **Worksheet** and then click **OK**.

In the new worksheet:

3. Enter a title in cell **A1**, **Bins** in cell **A3**, **Frequency** in cell **B3**, and **Midpoints** in cell **C3**.
4. Copy the bin number list in the cell range **D2:D12** of the **UNSTACKED worksheet** and paste this list into cell **A4** of the new worksheet.
5. Enter '--' in cell **C4**. Copy the midpoints list in the cell range **E2:E11** of the **UNSTACKED worksheet** and paste this list into cell **C5** of the new worksheet.
6. Select the cell range **B4:B14** that will hold the array formula.
7. Type, but do not press the **Enter** or **Tab** key, the formula **=FREQUENCY(UNSTACKED!\$A\$2:\$A\$228, \$A\$4: \$A\$14)**. Then, while holding down the **Ctrl** and **Shift** keys, press the **Enter** key to enter the array formula into the cell range **B4:B14**.
8. Adjust the worksheet formatting as necessary.

Steps 1 through 8 construct a frequency distribution for the growth retirement funds. To construct a frequency distribution for the value retirement funds, repeat steps 1 through 8 but in step 7 type **=FREQUENCY(UNSTACKED!\$B\$1:\$B\$90, \$A\$4: \$A\$14)** as the array formula.

Having constructed the two frequency distributions, continue by constructing the two histograms. Open to the worksheet that contains the frequency distribution for the growth funds and:

1. Select the cell range **B3:B14** (the cell range of the frequencies).
2. In Excel 2013, select **Insert**, then the **Column icon** in the **Charts group** (#3 in the illustration on page 89), and then select the **first 2-D Column** gallery item (**Clustered Column**). In other Excels, select **Insert → Column** and select the **first 2-D Column** gallery item (**Clustered Column**).
3. Right-click the chart and click **Select Data** in the shortcut menu.

In the **Select Data Source** dialog box:

4. Click **Edit** under the **Horizontal (Categories) Axis Labels** heading.
5. In the **Axis Labels** dialog box, drag the mouse to *select* the cell range **C4:C14** (containing the midpoints) to enter that

cell range. Do not type this cell range in the **Axis label range** box as you would otherwise do for the reasons explained in Appendix Section B.7. Click **OK** in this dialog box and then click **OK** (in the **Select Data Source** dialog box).

In the chart:

6. Right-click inside a bar and click **Format Data Series** in the shortcut menu.
7. (Excel 2013) In the **Format Data Series** task pane, click **Series Options**. In the **Series Options**, click **Series Options**, enter **0** in the **Gap Width** box, and then close the task pane. (To see the **Series Options**, you may have to first click the chart [third] icon near the top of the task pane.)
- (Other Excels) In the **Format Data Series** dialog box, click **Series Options** in the left pane, and in the **Series Options** right pane, change the **Gap Width** slider to **No Gap**. Click **Close**.
8. Relocate the chart to a chart sheet, turn off the chart legend and gridlines, add axis titles, and modify the chart title by using the instructions in Appendix Section B.6.

This example uses the workaround discussed in Section 2.2, “Classes and Excel Bins” on page 49. When you use this workaround, the histogram bar labeled—will *always* be a zero bar. Appendix Section B.8 explains how you can delete this unnecessary bar from the histogram, as was done for the examples shown in Section 2.4.

Analysis ToolPak

Use **Histogram**. For the example, open to the **UNSTACKED worksheet** of the **Retirement Funds workbook** and:

1. Select **Data → Data Analysis**. In the **Data Analysis** dialog box, select **Histogram** from the **Analysis Tools** list and then click **OK**.

In the **Histogram** dialog box:

2. Enter **A1:A228** as the **Input Range** and enter **D1:D12** as the **Bin Range**.
3. Check **Labels**, click **New Worksheet Ply**, and check **Chart Output**.
4. Click **OK** to create the frequency distribution and histogram on a new worksheet.

In the new worksheet:

5. Follow steps 5 through 9 of the **Analysis ToolPak** instructions in “The Frequency Distribution” in Section EG2.2.

These steps construct a frequency distribution and histogram for the growth funds. To construct a frequency distribution and histogram for the value funds, repeat the nine steps but in step 2 enter **B1:B90** as the **Input Range**. You will need to correct several formatting errors that Excel makes to the histograms it constructs. For each histogram:

1. Right-click inside a bar and click **Format Data Series** in the shortcut menu.
2. (Excel 2013) In the **Format Data Series** task pane, click **Series Options**. In the **Series Options**, click **Series Options**, enter **0** in the **Gap Width** box, and then close the task pane. (To see the **Series Options**, you may have to first click the chart [third] icon near the top of the task pane.)

(Other Excels) In the Format Data Series dialog box, click **Series Options** in the left pane, and in the Series Options right pane, change the **Gap Width** slider to **No Gap**. Click **Close**.

Histogram bars are labeled by bin numbers. To change the labeling to midpoints, open to each of the new worksheets and:

3. Enter **Midpoints** in cell **C1** and '--' in cell **C2**. Copy the cell range **E2:E11** of the **UNSTACKED worksheet** and paste this list into cell **C5** of the new worksheet.
4. Right-click the histogram and click **Select Data**.
5. In the Select Data Source dialog box, click **Edit** under the **Horizontal (Categories) Axis Labels** heading.
6. In the Axis Labels dialog box, drag the mouse to select the cell range **C2:C12** to enter that cell range. Do not type this cell range in the Axis label range box as you would otherwise do for the reasons explained in Appendix Section B.7. Click **OK** in this dialog box and then click **OK** (in the Select Data Source dialog box).
7. Relocate the chart to a chart sheet, turn off the chart legend and modify the chart title by using the instructions in Appendix Section B.6.

This example uses the workaround discussed in Section 2.2, "Classes and Excel Bins." Appendix Section B.8 explains how you can delete this unnecessary bar from the histogram, as was done for the examples shown in Section 2.4.

The Percentage Polygon and the Cumulative Percentage Polygon (Ogive)

Key Technique Modify an Excel line chart that is based on a frequency distribution.

Example Construct percentage polygons and cumulative percentage polygons for the one-year return percentages for the growth and value retirement funds, similar to Figure 2.11 on page 61 and equivalent to Figure 2.12 on page 62.

PHStat Use Histogram & Polygons.

For the example, use the **PHStat** instructions for creating a histogram on page 91 but in step 3 of those instructions, also check **Percentage Polygon** and **Cumulative Percentage Polygon (Ogive)** before clicking **OK**.

In-Depth Excel Use the Polygons workbook as a model.

For the example, open to the **UNSTACKED worksheet** of the **Retirement Funds workbook** and follow steps 1 through 8 to construct a frequency distribution for the growth retirement funds. Repeat steps 1 through 8 but in step 7 type the array formula **=FREQUENCY(UNSTACKED!\$B\$1:\$B\$90, \$A\$4: \$A\$14)** to construct a frequency distribution for the value funds. Open to the worksheet that contains the growth funds frequency distribution and:

1. Select column **C**. Right-click and click **Insert** in the shortcut menu. Right-click and click **Insert** in the shortcut menu a second time. (The worksheet contains new, blank columns C and D and the midpoints column is now column E.)
2. Enter **Percentage** in cell **C3** and **Cumulative Percentage** in cell **D3**.
3. Enter **=B4/SUM(\$B\$4:\$B\$14)** in cell **C4** and copy this formula down through **row 14**.

4. Enter **= C4** in cell **D4**.
5. Enter **= C5 + D4** in cell **D5** and copy this formula down through row **14**.
6. Select the cell range **C4:D14**, right-click, and click **Format Cells** in the shortcut menu.
7. In the **Number** tab of the Format Cells dialog box, click **Percentage** in the **Category** list and click **OK**.

Open to the worksheet that contains the value funds frequency distribution and repeat steps 1 through 7. To construct the percentage polygons, open to the worksheet that contains the growth funds distribution and:

1. Select cell range **C4:C14**.
2. In Excel 2013, select **Insert**, then select the **Line icon** in the **Charts group** (#4 in the illustration on page 89), and then select the **fourth 2-D Line** gallery item (**Line with Markers**). In other Excels, select **Insert → Line** and select the **fourth 2-D Line** gallery item (**Line with Markers**).
3. Right-click the chart and click **Select Data** in the shortcut menu.

In the Select Data Source dialog box:

4. Click **Edit** under the **Legend Entries (Series)** heading. In the Edit Series dialog box, enter the *formula* **= "Growth Funds"** as the **Series name** and click **OK**.
5. Click **Edit** under the **Horizontal (Categories) Axis Labels** heading. In the Axis Labels dialog box, drag the mouse to select the cell range **E4:E14** to enter that cell range. Do not type this cell range in the Axis label range box as you would otherwise do for the reasons explained in Appendix Section B.7.
6. Click **OK** in this dialog box and then click **OK** (in the Select Data Source dialog box).

Back in the chart:

7. Relocate the chart to a chart sheet, turn off the chart gridlines, add axis titles, and modify the chart title by using the instructions in Appendix Section B.6.

In the new chart sheet:

8. Right-click the chart and click **Select Data** in the shortcut menu.
9. In the Select Data Source dialog box, click **Add**.

In the Edit Series dialog box:

10. Enter the *formula* **= "Value Funds"** as the **Series name** and press **Tab**.
11. With the current value in **Series values** highlighted, click the worksheet tab for the worksheet that contains the value funds distribution.
12. Drag the mouse to select the cell range **C4:C14** to enter that cell range as the **Series values**. Do not type this cell range in the Series values box as you would otherwise do, for the reasons explained in Appendix Section B.7.
13. Click **OK**. Back in the Select Data Source dialog box, click **OK**.

To construct the cumulative percentage polygons, open to the worksheet that contains the growth funds distribution and repeat steps 1 through 13 but replace steps 1, 5, and 12 with these steps:

1. Select the cell range **D4:D14**.
5. Click **Edit** under the **Horizontal (Categories) Axis Labels** heading. In the Axis Labels dialog box, drag the mouse to select the cell range **A4:A14** to enter that cell range.
12. Drag the mouse to select the cell range **D4:D14** to enter that cell range as the **Series values**.

If the *Y* axis of the cumulative percentage polygon extends past 100%, right-click the axis and click **Format Axis** in the shortcut menu. In the Excel 2013 Format Axis task pane, click **Axis Options**. In the Axis Options, enter **0** in the **Minimum** box and then close the pane. In other Excels, you set this value in the Format Axis dialog box. Click **Axis Options** in the left pane, and in the Axis Options right pane, click the first **Fixed** option button (for Minimum), enter **0** in its box, and then click **Close**.

EG2.5 VISUALIZING TWO NUMERICAL VARIABLES

The Scatter Plot

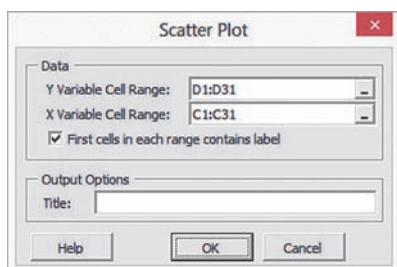
Key Technique Use the Excel scatter chart.

Example Construct a scatter plot of revenue and value for NBA teams, similar to Figure 2.14 on page 66.

PHStat Use Scatter Plot.

For the example, open to the **DATA worksheet** of the **NBAValues workbook**. Select **PHStat → Descriptive Statistics → Scatter Plot**. In the procedure's dialog box (shown below):

1. Enter **D1:D31** as the **Y Variable Cell Range**.
2. Enter **C1:C31** as the **X Variable Cell Range**.
3. Check **First cells in each range contains label**.
4. Enter a **Title** and click **OK**.



To add a superimposed line like the one shown in Figure 2.14, click the chart and use step 3 of the *In-Depth Excel* instructions.

In-Depth Excel Use the **Scatter Plot workbook** as a model. For the example, open to the **DATA worksheet** of the **NBAValues workbook** and:

1. Select the cell range **C1:D31**.
2. In Excel 2013, select **Insert**, then the **Scatter (X,Y)** icon in the **Charts group** (#5 in the illustration on page 89), and then select the **first Scatter** gallery item (**Scatter**). In other Excels,

select **Insert → Scatter** and select the **first Scatter** gallery item (**Scatter with only Markers**).

3. In Excel 2013, select **Design → Add Chart Element → Trendline → Linear**. In other Excels, select **Layout → Trendline → Linear Trendline**.
4. Relocate the chart to a chart sheet, turn off the chart legend and gridlines, add axis titles, and modify the chart title by using the instructions in Appendix Section B.6.

When constructing Excel scatter charts with other variables, make sure that the *X* variable column precedes (is to the left of) the *Y* variable column. (If the worksheet is arranged *Y* then *X*, cut and paste so that the *Y* variable column appears to the right of the *X* variable column.)

The Time-Series Plot

Key Technique Use the Excel scatter chart.

Example Construct a time-series plot of movie revenue per year from 1995 to 2012, similar to Figure 2.15 on page 67.

In-Depth Excel Use the **Time Series workbook** as a model. For the example, open to the **DATA worksheet** of the **Movie Revenues workbook** and:

1. Select the cell range **A1:B19**.
2. In Excel 2013, select **Insert**, then select the **Scatter (X, Y)** icon in the **Charts group** (#5 in the illustration on page 89), and then select the **fourth Scatter** gallery item (**Scatter with Straight Lines and Markers**). In other Excels, select **Insert → Scatter** and select the **fourth Scatter** gallery item (**Scatter with Straight Lines and Markers**).
3. Relocate the chart to a chart sheet, turn off the chart legend and gridlines, add axis titles, and modify the chart title by using the instructions in Appendix Section B.6.

When constructing time-series charts with other variables, make sure that the *X* variable column precedes (is to the left of) the *Y* variable column. (If the worksheet is arranged *Y* then *X*, cut and paste so that the *Y* variable column appears to the right of the *X* variable column.)

EG2.6 ORGANIZING MANY CATEGORICAL VARIABLES

Multidimensional Contingency Tables

Key Technique Use the Excel PivotTable feature.

Example Construct a PivotTable showing percentage of overall total for fund type, risk, and market cap for the retirement funds sample, similar to the one shown at the right in Figure 2.16 on page 69.

In-Depth Excel Use the **MCT workbook** as a model. For the example, open to the **DATA worksheet** of the **Retirement Funds workbook** and:

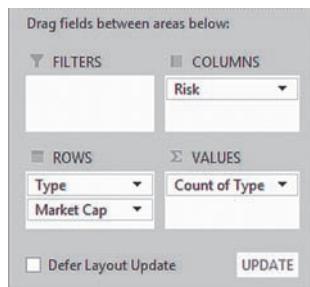
1. Select **Insert → PivotTable**.

In the Create PivotTable dialog box:

2. Click **Select a table or range** and enter A1:N317 as the **Table/Range**.
3. Click **New Worksheet** and then click **OK**.

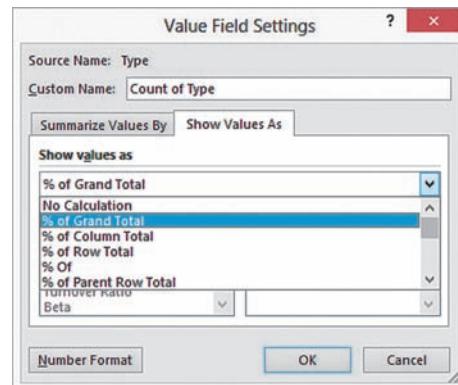
Excel inserts a new worksheet and displays the PivotTable Field List pane. The worksheet contains a graphical representation of a PivotTable that will change as you work inside the PivotTable Field List (or PivotTable Fields) task pane. In that pane (partially shown in the next column):

4. Drag **Type** in the **Choose fields to add to report** box and drop it in the **ROWS** (or **Row Labels**) box.
5. Drag **Market Cap** in the **Choose fields to add to report** box and drop it in the **ROWS** (or **Row Labels**) box.
6. Drag **Risk** in the **Choose fields to add to report** box and drop it in the **COLUMNS** (or **Column Labels**) box.
7. Drag **Type** in the **Choose fields to add to report** box a second time and drop it in the **Σ VALUES** box. The dropped label changes to **Count of Type**.
8. Click (not right-click) the dropped label **Count of Type** and click **Value Field Settings** in the shortcut menu.



In the Value Field Settings dialog box:

9. Click the **Show Values As** tab and select **% of Grand Total** from the **Show values as** drop-down list (shown below).
10. Click **OK**.



In the PivotTable:

11. Enter a title in cell A1.
12. Enter a **space character** in cell A3 to replace the value "Count of Type."
13. Follow steps 8 and 9 of the *In-Depth Excel* "The Contingency Table" instructions on page 89 to relocate the Low column from column D to column B.

If the PivotTable you construct does not contain a row and column for the grand totals as the PivotTables in Figure 2.21 contain, follow steps 10 through 15 of the *In-Depth Excel*, "The Contingency Table" instructions to include the grand totals.

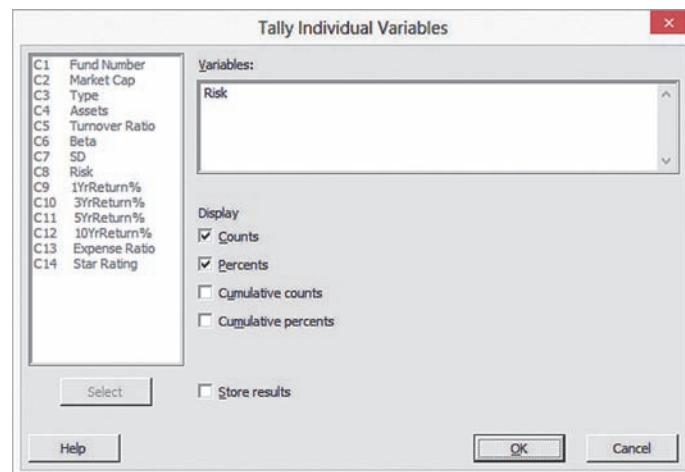
CHAPTER 2 MINITAB GUIDE

MG2.1 ORGANIZING CATEGORICAL VARIABLES

The Summary Table

Use **Tally Individual Variables** to create a summary table. For example, to create a summary table similar to Table 2.3 on page 38, open to the **Retirement Funds worksheet**. Select **Stat → Tables → Tally Individual Variables**. In the procedure's dialog box (shown at right):

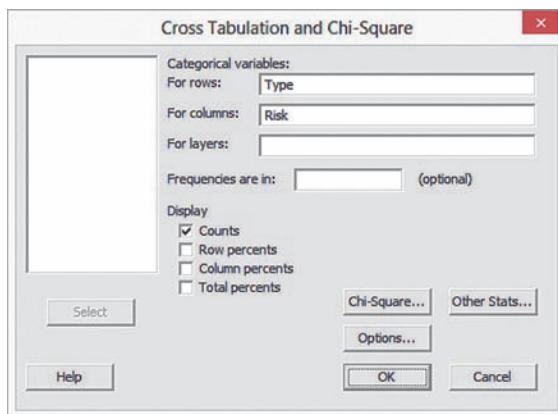
1. Double-click **C8 Risk** in the variables list to add **Risk** to the **Variables** box.
2. Check **Counts** and **Percents**.
3. Click **OK**.



The Contingency Table

Use **Cross Tabulation and Chi-Square** to create a contingency table. For example, to create a contingency table similar to Table 2.4 on page 39, open to the **Retirement Funds worksheet**. Select **Stat → Tables → Cross Tabulation and Chi-Square**. In the procedure's dialog box (shown below):

1. Enter **Type** in the **For rows** box.
2. Enter **Risk** in the **For columns** box
3. Check **Counts**.
4. Click **OK**.



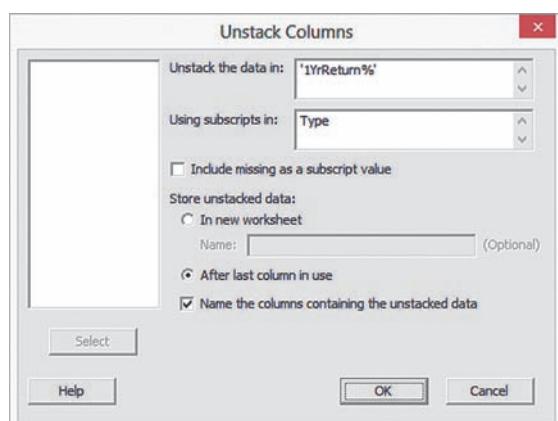
To create the other types of contingency tables shown in Tables 2.5 through 2.7, check **Row percents**, **Column percents**, or **Total percents**, respectively, in step 3.

MG2.2 ORGANIZING NUMERICAL VARIABLES

Stacked and Unstacked Data

Use **Stack** or **Unstack Columns** to rearrange data. For example, to unstack the **1YrReturn%** variable in column C9 of the **Retirement Funds worksheet** by fund type, open to that worksheet. Select **Data → Unstack Columns**. In the procedure's dialog box (shown below):

1. Double-click **C9 1YrReturn%** in the variables list to add '**1YrReturn%**' to the **Unstack the data in** box and press **Tab**.
2. Double-click **C3 Type** in the variables list to add **Type** to the **Using subscripts in** box.
3. Click **After last column in use**.
4. Check **Name the columns containing the unstacked data**.
5. Check **OK**.



Minitab inserts two new columns, **1YrReturn%_Growth** and **1YrReturn%_Value**, the names of which you can edit.

To stack columns, select **Data → Stack → Columns**. In the **Stack Columns** dialog box, add the names of columns that contain the data to be stacked to the **Stack the following columns** box and then click either **New worksheet** or **Column of current worksheet** as the place to store the stacked data.

The Ordered Array

Use **Sort** to create an ordered array. Select **Data → Sort** and in the **Sort** dialog box (not shown), double-click a column name in the variables list to add it to the **Sort column(s)** box and then press **Tab**. Double-click the same column name in the variables list to add it to the first **By column** box. Click either **New worksheet**, **Original column(s)**, or **Column(s) of current worksheet**. (If you choose the third option, also enter the name of the column in which to place the ordered data in the box). Click **OK**.

The Frequency Distribution

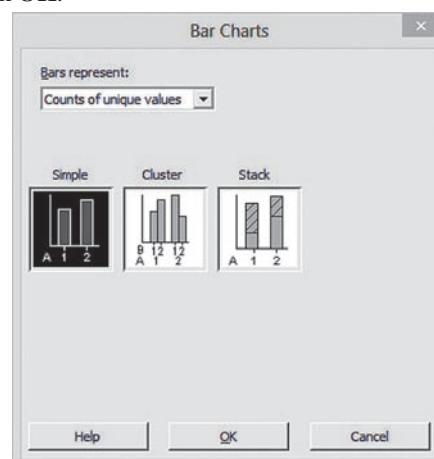
There are no Minitab commands that use classes that you specify to create frequency distributions of the type seen in Tables 2.9 through 2.12. (See also “The Histogram” in Section MG2.4.)

MG2.3 VISUALIZING CATEGORICAL VARIABLES

The Bar Chart and the Pie Chart

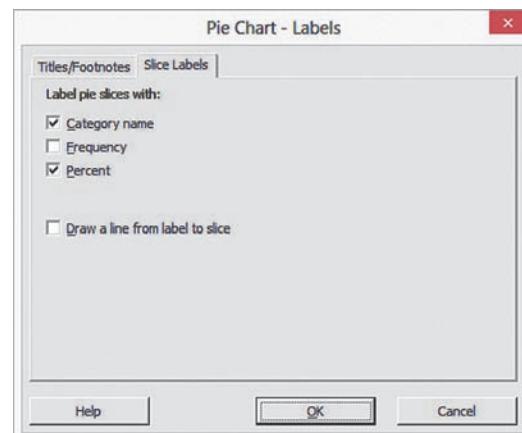
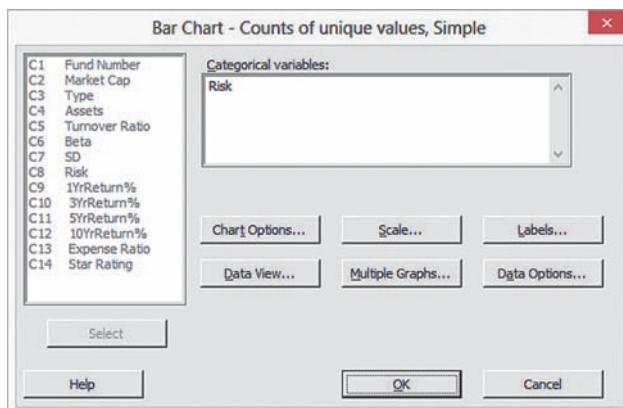
Use **Bar Chart** to create a bar chart from a summary table and use **Pie Chart** to create a pie chart from a summary table. For example, to create the Figure 2.2 bar chart on page 38, open to the **Retirement Funds worksheet**. Select **Graph → Bar Chart**. In the procedure's dialog box (shown below):

1. Select **Counts of unique values** from the **Bars represent** drop-down list.
2. In the gallery of choices, click **Simple**.
3. Click **OK**.



In the Bar Chart - Counts of unique values, Simple dialog box (shown at top of left column on next page):

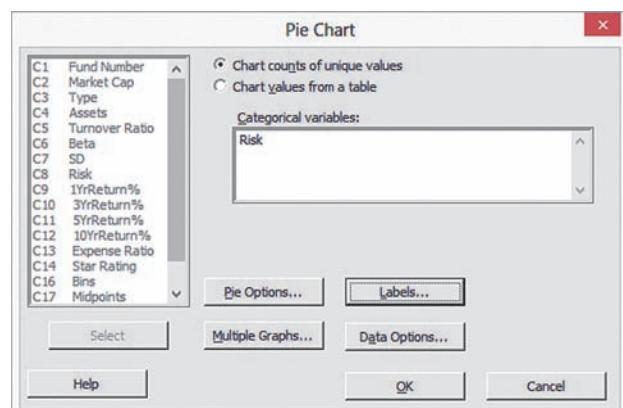
4. Double-click **C8 Risk** in the variables list to add **Risk** to **Categorical variables**.
5. Click **OK**.



If your data are in the form of a table of frequencies, select **Values from a table** from the **Bars represent** drop-down list in step 1. With this selection, clicking **OK** in step 3 will display the “Bar Chart - Values from a table, One column of values, Simple” dialog box. In this dialog box, you enter the columns to be graphed in the **Graph variables** box and, optionally, enter the column in the worksheet that holds the categories for the table in the **Categorical variable** box.

Use **Pie Chart** to create a pie chart from a summary table. For example, to create the Figure 2.3 pie chart on page 38, open to the **Retirement Funds worksheet**. Select **Graph → Pie Chart**. In the Pie Chart dialog box (shown below):

1. Click **Chart counts of unique values** and then press **Tab**.
2. Double-click **C8 Risk** in the variables list to add **Risk** to **Categorical variables**.
3. Click **Labels**.



In the Pie Chart - Labels dialog box (shown at top of next column):

4. Click the **Slice Labels** tab.
5. Check **Category name** and **Percent**.
6. Click **OK** to return to the original dialog box.

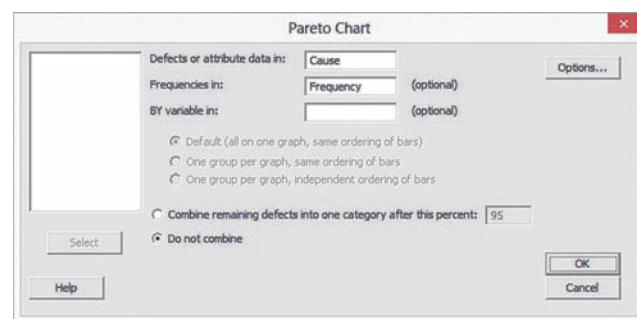
Back in the original Pie Chart dialog box:

7. Click **OK**.

The Pareto Chart

Use **Pareto Chart** to create a Pareto chart. For example, to create the Figure 2.4 Pareto chart on page 54, open to the **ATM Transactions worksheet**. Select **Stat → Quality Tools → Pareto Chart**. In the procedure’s dialog box (shown below):

1. Double-click **C1 Cause** in the variables list to add **Cause** to the **Defects or attribute data in** box.
2. Double-click **C2 Frequency** in the variables list to add **Frequency** to the **Frequencies in** box.
3. Click **Do not combine**.
4. Click **OK**.



The Side-by-Side Chart

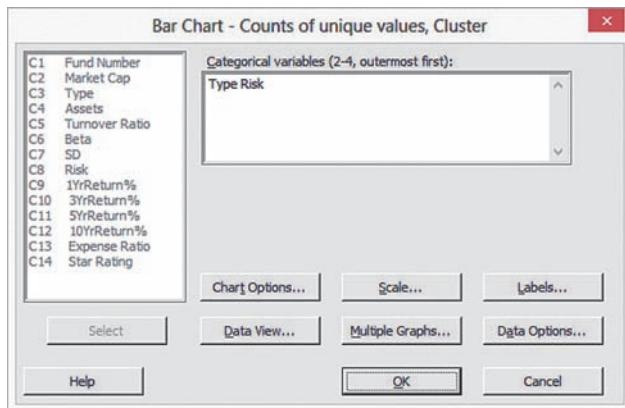
Use **Bar Chart** to create a side-by-side chart. For example, to create the Figure 2.6 side-by-side chart on page 55, open to the **Retirement Funds worksheet**. Select **Graph → Bar Chart**. In the Bar Charts dialog box:

1. Select **Counts of unique values** from the **Bars represent** drop-down list.
2. In the gallery of choices, click **Cluster**.
3. Click **OK**.

In the “Bar Chart - Counts of unique values, Cluster” dialog box (shown below):

4. Double-click **C3 Type** and **C8 Risk** in the variables list to add **Type** and **Risk** to the **Categorical variables (2–4, outermost first)** box.

5. Click **OK**.

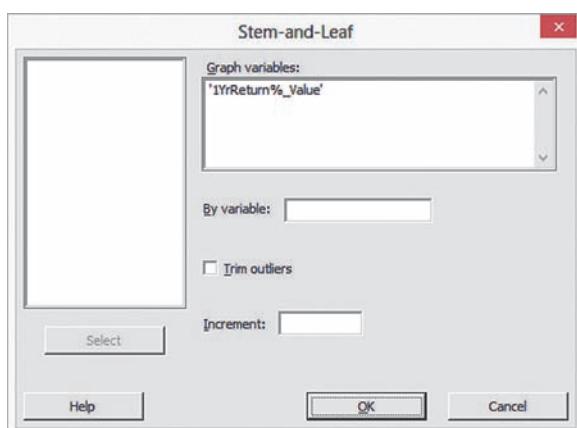


MG2.4 VISUALIZING NUMERICAL VARIABLES

The Stem-and-Leaf Display

Use **Stem-and-Leaf** to create a stem-and-leaf display. For example, to create the Figure 2.7 stem-and-leaf display on page 58, open to the **Unstacked1YrReturn Funds** worksheet. Select **Graph → Stem-and-Leaf**. In the procedure’s dialog box (shown below):

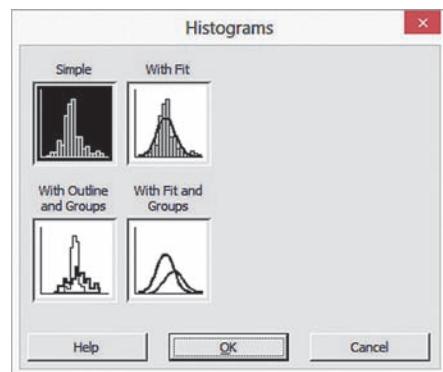
1. Double-click **C2 1YrReturn%_Value** in the variables list to add '**1YrReturn%_Value**' in the **Graph variables** box.
2. Click **OK**.



The Histogram

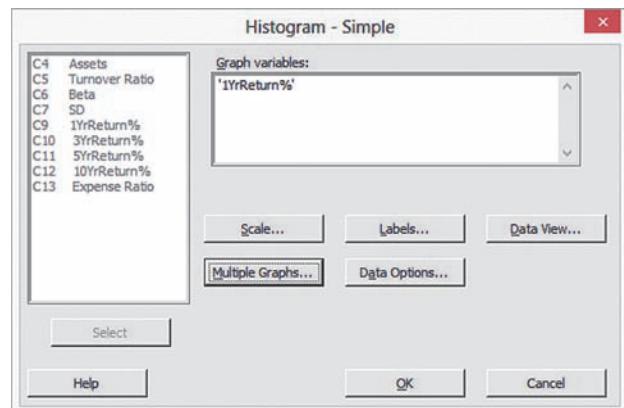
Use **Histogram** to create a histogram. For example, to create the pair of histograms shown in Figure 2.9 on page 60, open to the **Retirement Funds** worksheet. Select **Graph → Histogram**. In the Histograms dialog box (shown at top of next column):

1. Click **Simple** and then click **OK**.



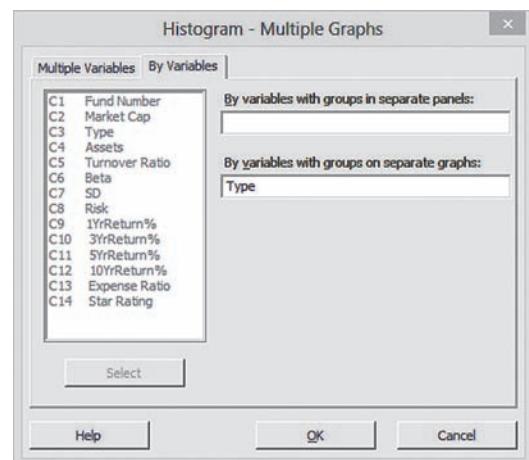
In the Histogram - Simple dialog box (shown below):

2. Double-click **C9 1YrReturn%** in the variables list to add '**1YrReturn%**' in the **Graph variables** box.
3. Click **Multiple Graphs**.



In the Histogram - Multiple Graphs dialog box:

4. In the **Multiple Variables** tab (not shown), click **On separate graphs** and then click the **By Variables** tab.
5. In the **By Variables** tab (shown below), enter **Type** in the **By variables in groups on separate graphs** box.
6. Click **OK**.



Back in the Histogram - Simple dialog box:

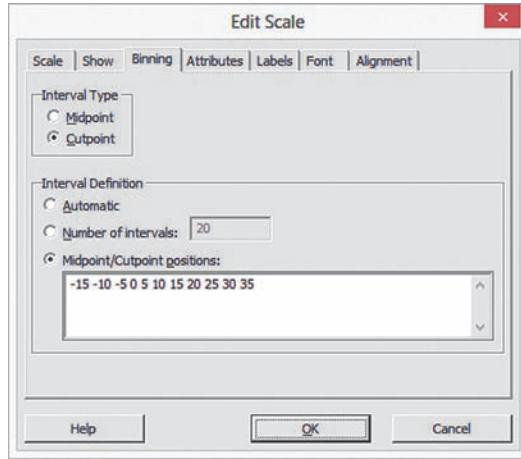
7. Click OK.

The histograms created use classes that differ from the classes used in Figure 2.9 (and in Table 2.10 on page 44) and do not use the midpoints shown in Figure 2.9. To better match the histograms shown in Figure 2.9, for each histogram:

8. Right-click the X axis and then click **Edit X Scale from the shortcut menu.**

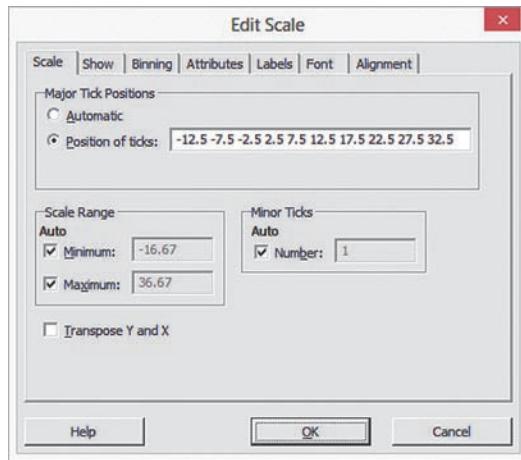
In the Edit Scale dialog box:

9. Click the **Binning tab (shown below). Click **Cutpoint** (as the **Interval Type**) and **Midpoint/Cutpoint positions** and enter **-15 -10 -5 0 5 10 15 20 25 30 35** in the box (with a space after each value).**



10. Click the **Scale tab (shown below). Click **Position of ticks** and enter **-12.5 -7.5 -2.5 2.5 7.5 12.5 17.5 22.5 27.5 32.5** in the box (with a space after each value).**

11. Click OK.



To create the histogram of the one-year return percentage variable for all funds in the retirement fund sample, repeat steps 1 through 11, but in step 5 delete **Type** from the **By variables in groups on separate graphs** box.

To modify the histogram bars, double-click over the histogram bars and make the appropriate entries and selections in the Edit Bars dialog box. To modify an axis, double-click the axis and make the appropriate entries and selections in the Edit Scale dialog box.

The Percentage Polygon

Use **Histogram** to create a percentage polygon. For example, to create the pair of percentage polygons shown in Figure 2.11 on page 61, open to the **Unstacked 1YrReturn** worksheet. Select **Graph → Histogram**. In the Histograms dialog box:

1. Click **Simple and then click **OK**.**

In the Histogram - Simple dialog box:

- 2. Double-click **C1 1YrReturn%_Growth** in the variables list to add '**1YrReturn%_Growth**' in the **Graph variables** box.**
- 3. Double-click **C2 1YrReturn%_Value** in the variables list to add '**1YrReturn%_Value**' in the **Graph variables** box.**
- 4. Click **Scale**.**

In the Histogram - Scale dialog box:

- 5. Click the **Y-Scale Type** tab. Click **Percent**, clear **Accumulate values across bins**, and then click **OK**.**

Back again in the Histogram - Simple dialog box:

6. Click **Data View.**

In the Histogram - Data View dialog box:

- 7. Click the **Data Display** tab. Check **Symbols** and clear all of the other check boxes.**
- 8. Click the **Smoother** tab and then click **Lowness** and enter **0** as the **Degree of smoothing** and **1** as the **Number of steps**.**
- 9. Click **OK**.**

Back again in the Histogram - Simple dialog box:

10. Click **OK to create the polygons.**

The percentage polygons created do not use the classes and midpoints shown in Figure 2.11. To better match the polygons shown in Figure 2.11:

- 11. Right-click the X axis and then click **Edit X Scale** from the shortcut menu.**

In the Edit Scale dialog box:

- 12. Click the **Binning** tab. Click **Cutpoint** as the **Interval Type** and **Midpoint/Cutpoint positions** and enter **-15 -10 -5 0 5 10 15 20 25 30 35** in the box (with a space after each value).**
- 13. Click the **Scale** tab. Click **Position of ticks** and enter **-12.5 -7.5 -2.5 2.5 7.5 12.5 17.5 22.5 27.5 32.5** in the box (with a space after each value).**
- 14. Click **OK**.**

The Cumulative Percentage Polygon (Ogive)

Modify the “The Percentage Polygon” instructions to create a cumulative percentage polygon. Replace steps 5 and 12 with the following steps:

- 5. Click the **Y-Scale Type** tab. Click **Percent**, check **Accumulate values across bins**, and then click **OK**.**
- 12. Click the **Binning** tab. Click **Midpoint** as the **Interval Type** and **Midpoint/Cutpoint positions** and enter **-15 -10 -5 0 5 10 15 20 25 30 35** in the box (with a space after each value).**

MG2.5 VISUALIZING TWO NUMERICAL VARIABLES

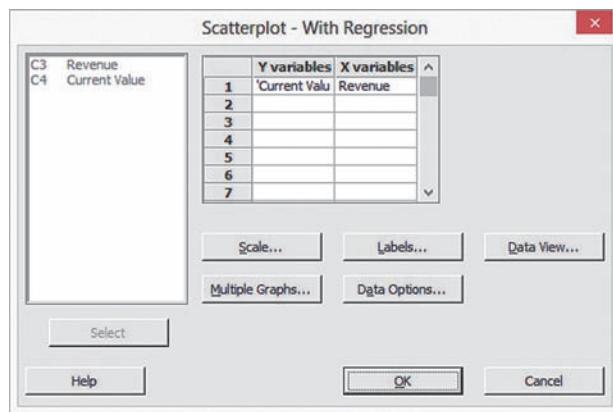
The Scatter Plot

Use **Scatterplot** to create a scatter plot. For example, to create a scatter plot similar to the one shown in Figure 2.14 on page 66, open to the **NBAValues worksheet**. Select **Graph → Scatterplot**. In the Scatterplots dialog box:

1. Click **With Regression** and then click **OK**.

In the Scatterplot - With Regression dialog box (shown below):

2. Double-click **C4 Current Value** in the variables list to enter 'Current Value' in the **row 1 Y variables** cell.
3. Enter **Revenue** in the **row 1 X variables** cell.
4. Click **OK**.



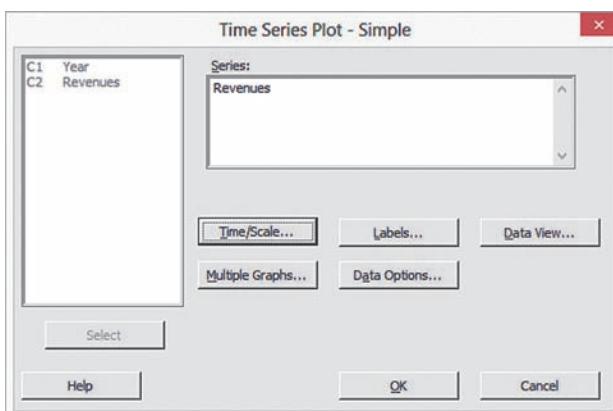
The Time-Series Plot

Use **Time Series Plot** to create a time-series plot. For example, to create the Figure 2.15 time-series plot on page 67, open to the **Movie Revenues worksheet** and select **Graph → Time Series Plot**. In the Time Series Plots dialog box:

1. Click **Simple** and then click **OK**.

In the Time Series Plot - Simple dialog box (shown below):

2. Double-click **C2 Revenues** in the variables list to add **Revenues** in the **Series** box.
3. Click **Time/Scale**.

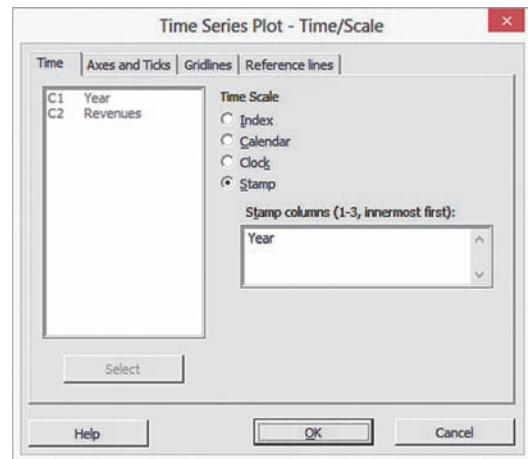


In the Time Series Plot - Time/Scale dialog box (shown below):

4. Click **Stamp** and then press **Tab**.
5. Double-click **C1 Year** in the variables list to add **Year** in the **Stamp columns (1-3, innermost first)** box.
6. Click **OK**.

Back in the Time Series Plot - Simple dialog box:

7. Click **OK**.



MG2.6 ORGANIZING MANY CATEGORICAL VARIABLES

Multidimensional Contingency Tables

Use **Cross Tabulation and Chi-Square** to create a multidimensional contingency table. For example, to create a table similar to the Figure 2.16 fund type, market cap, and risk table on page 69, open to the **Retirement Funds worksheet**. Select **Stat → Tables → Cross Tabulation and Chi-Square**. In the procedure's dialog box:

1. Double-click **C3 Type** in the variables list to add **Type** to the **For rows** box.
2. Double-click **C2 Market Cap** in the variables list to add '**Market Cap**' to the **For rows** box and then press **Tab**.
3. Double-click **C8 Risk** in the variables list to add **Risk** to the **For columns** box.
4. Check **Counts**.
5. Click **OK**.

To display the cell values as percentages, as was done in Figure 2.1, check **Total percents** instead of **Counts** in step 4.

CHAPTER 3

Numerical Descriptive Measures

CONTENTS

- 3.1 Central Tendency
- 3.2 Variation and Shape

VISUAL EXPLORATIONS: Exploring Descriptive Statistics

- 3.3 Exploring Numerical Data
- 3.4 Numerical Descriptive Measures for a Population
- 3.5 The Covariance and the Coefficient of Correlation
- 3.6 Descriptive Statistics: Pitfalls and Ethical Issues

USING STATISTICS: More Descriptive Choices, Revisited

CHAPTER 3 EXCEL GUIDE

CHAPTER 3 MINITAB GUIDE

OBJECTIVES

- To describe the properties of central tendency, variation, and shape in numerical data
- To construct and interpret a boxplot
- To compute descriptive summary measures for a population
- To compute the covariance and the coefficient of correlation

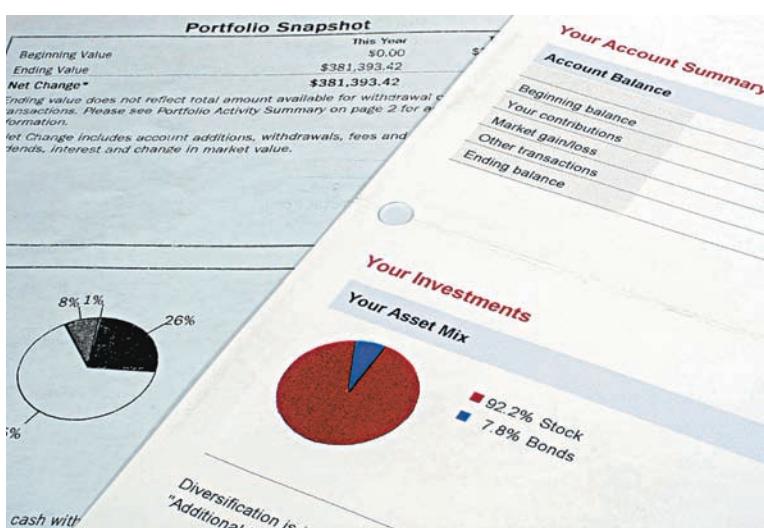
USING STATISTICS

More Descriptive Choices

As a member of a Choice *Is Yours* investment service task force, you helped organize and visualize the variables found in a sample of 316 retirement funds. Now, several weeks later, prospective clients are asking for more information on which they can base their investment decisions. In particular, they would like to be able to compare the results of an individual retirement fund to the results of similar funds.

For example, while the earlier work your team did shows how the one-year return percentages are distributed, prospective clients would like to know how the value for a particular mid-cap growth fund compares to the one-year returns of all mid-cap growth funds. They also seek to understand how the values for the variables collected vary. Are all the values relatively similar? And does any variable have outlier values that are either extremely small or extremely large?

While doing a complete search of the retirement funds data could lead to answers to the preceding questions, you wonder if there are easier ways than extensive searching to uncover those answers. You also wonder if there are other ways of being more *descriptive* about the sample of funds—providing answers to questions not yet raised by prospective clients. If you can help the Choice *Is Yours* investment service provide such answers, prospective clients will be better able to evaluate the retirement funds that your firm features.



Baranq/Shutterstock

The prospective clients in the More Descriptive Choices scenario have begun asking questions about numerical variables such as the one-year return percentage. When summarizing and describing numerical variables, the organizing and visualizing methods discussed in Chapter 2 are only a starting point. You also need to describe such variables in terms of their central tendency, variation, and shape.

Central tendency is the extent to which the values of a numerical variable group around a typical, or central, value. **Variation** measures the amount of dispersion, or scattering, away from a central value that the values of a numerical variable show. The *shape* of a variable is the pattern of the distribution of values from the lowest value to the highest value.

This chapter discusses ways you can compute these numerical descriptive measures as you begin to analyze your data within the DCOVA framework. The chapter also talks about the covariance and the coefficient of correlation, measures that can help show the strength of the association between two numerical variables. Computing the descriptive measures discussed in this chapter would be one way to help prospective clients of the Choice *Is Yours* service find the answers they seek.

3.1 Central Tendency

Most variables show a distinct tendency to group around a central value. When people talk about an “average value” or the “middle value” or the “most frequent value,” they are talking informally about the mean, median, and mode—three measures of central tendency.

The Mean

The **arithmetic mean** (typically referred to as the **mean**) is the most common measure of central tendency. The mean can suggest a typical or central value and serves as a “balance point” in a set of data, similar to the fulcrum on a seesaw. The mean is the only common measure in which all the values play an equal role. You compute the mean by adding together all the values and then dividing that sum by the number of values in the data set.

The symbol \bar{X} , called *X-bar*, is used to represent the mean of a sample. For a sample containing n values, the equation for the mean of a sample is written as

$$\bar{X} = \frac{\text{sum of the values}}{\text{number of values}}$$

Using the series X_1, X_2, \dots, X_n to represent the set of n values and n to represent the number of values in the sample, the equation becomes

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

By using summation notation (discussed fully in Appendix A), you replace the numerator $X_1 + X_2 + \dots + X_n$ with the term $\sum_{i=1}^n X_i$, which means sum all the X_i values from the first X value, X_1 , to the last X value, X_n , to form Equation (3.1), a formal definition of the sample mean.

SAMPLE MEAN

The **sample mean** is the sum of the values in a sample divided by the number of values in the sample:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (3.1)$$

where

\bar{X} = sample mean

n = number of values or sample size

X_i = i th value of the variable X

$$\sum_{i=1}^n X_i = \text{summation of all } X_i \text{ values in the sample}$$

Because all the values play an equal role, a mean is greatly affected by any value that is very different from the others. When you have such extreme values, you should avoid using the mean as a measure of central tendency.

For example, if you knew the typical time it takes you to get ready in the morning, you might be able to arrive at your first destination every day in a more timely manner. Using the DCOVA framework, you first define the time to get ready as the time from when you get out of bed to when you leave your home, rounded to the nearest minute. Then, you collect the times for 10 consecutive workdays and organize and store them in **Times**.

Using the collected data, you compute the mean to discover the “typical” time it takes for you to get ready. For these data:

Day:	1	2	3	4	5	6	7	8	9	10
Time (minutes):	39	29	43	52	39	44	40	31	44	35

the mean time is 39.6 minutes, computed as follows:

$$\bar{X} = \frac{\text{sum of the values}}{\text{number of values}}$$

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$\begin{aligned}\bar{X} &= \frac{39 + 29 + 43 + 52 + 39 + 44 + 40 + 31 + 44 + 35}{10} \\ &= \frac{396}{10} = 39.6\end{aligned}$$

Even though no individual day in the sample had a value of 39.6 minutes, allotting this amount of time to get ready in the morning would be a reasonable decision to make. The mean is a good measure of central tendency in this case because the data set does not contain any exceptionally small or large values.

To illustrate how the mean can be greatly affected by any value that is very different from the others, imagine that on Day 3, a set of unusual circumstances delayed you getting ready by an extra hour, so that the time for that day was 103 minutes. This extreme value causes the mean to rise to 45.6 minutes, as follows:

$$\bar{X} = \frac{\text{sum of the values}}{\text{number of values}}$$

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$\bar{X} = \frac{39 + 29 + 103 + 52 + 39 + 44 + 40 + 31 + 44 + 35}{10}$$

$$\bar{X} = \frac{456}{10} = 45.6$$

The one extreme value has increased the mean by 6 minutes. The extreme value also moved the position of the mean relative to all the values. The original mean, 39.6 minutes, had a middle, or *central*, position among the data values: 5 of the times were less than that mean and 5 were greater than that mean. In contrast, the mean using the extreme value is greater than 9 of the 10 times, making the new mean a poor measure of central tendency.

EXAMPLE 3.1

The Mean Calories in Cereals

Nutritional data about a sample of seven breakfast cereals (stored in **Cereals**) includes the number of calories per serving:

Cereal	Calories
Kellogg's All Bran	80
Kellogg's Corn Flakes	100
Wheaties	100
Nature's Path Organic Multigrain Flakes	110
Kellogg's Rice Krispies	130
Post Shredded Wheat Vanilla Almond	190
Kellogg's Mini Wheats	200

Compute the mean number of calories in these breakfast cereals.

SOLUTION The mean number of calories is 130, computed as follows:

$$\bar{X} = \frac{\text{sum of the values}}{\text{number of values}}$$

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^n X_i}{n} \\ &= \frac{910}{7} = 130\end{aligned}$$

The Median

The **median** is the middle value in an ordered array of data that has been ranked from smallest to largest. Half the values are smaller than or equal to the median, and half the values are larger than or equal to the median. The median is not affected by extreme values, so you can use the median when extreme values are present.

To compute the median for a set of data, you first rank the values from smallest to largest and then use Equation (3.2) to compute the rank of the value that is the median.

MEDIAN

$$\text{Median} = \frac{n+1}{2} \text{ ranked value} \quad (3.2)$$

You compute the median by following one of two rules:

- **Rule 1** If the data set contains an *odd* number of values, the median is the measurement associated with the middle-ranked value.
- **Rule 2** If the data set contains an *even* number of values, the median is the measurement associated with the average of the two middle-ranked values.

To further analyze the sample of 10 times to get ready in the morning, you can compute the median. To do so, you rank the daily times as follows:

Student Tip

Remember that you must rank the values in order from the smallest to the largest to compute the median.

<i>Ranked values:</i>	29	31	35	39	39	40	43	44	44	52
<i>Ranks:</i>	1	2	3	4	5	6	7	8	9	10
↑										
Median = 39.5										

Because the result of dividing $n + 1$ by 2 for this sample of 10 is $(10 + 1)/2 = 5.5$, you must use Rule 2 and average the measurements associated with the fifth and sixth ranked values, 39 and 40. Therefore, the median is 39.5. The median of 39.5 means that for half the days, the time to get ready is less than or equal to 39.5 minutes, and for half the days, the time to get ready is greater than or equal to 39.5 minutes. In this case, the median time to get ready of 39.5 minutes is very close to the mean time to get ready of 39.6 minutes.

EXAMPLE 3.2

Computing the Median from an Odd-Sized Sample

Nutritional data about a sample of seven breakfast cereals (stored in **Cereals**) includes the number of calories per serving (see Example 3.1 on page 104). Compute the median number of calories in breakfast cereals.

SOLUTION Because the result of dividing $n + 1$ by 2 for this sample of seven is $(7 + 1)/2 = 4$, using Rule 1, the median is the measurement associated with the fourth-ranked value. The number of calories per serving values are ranked from the smallest to the largest:

<i>Ranked values:</i>	80	100	100	110	130	190	200
<i>Ranks:</i>	1	2	3	4	5	6	7
↑							
Median = 110							

The median number of calories is 110. Half the breakfast cereals have equal to or less than 110 calories per serving, and half the breakfast cereals have equal to or more than 110 calories.

The Mode

The **mode** is the value that appears most frequently. Like the median and unlike the mean, extreme values do not affect the mode. For a particular variable, there can be several modes or no mode at all. For example, for the sample of 10 times to get ready in the morning:

29 31 35 39 39 40 43 44 44 52

there are two modes, 39 minutes and 44 minutes, because each of these values occurs twice. However, for this sample of 7 smartphone prices offered by a cellphone provider (stored in **Smartphones**):

20 80 150 200 230 280 370

there is no mode. None of the values is “most typical” because each value appears the same number of times (once) in the data set.

EXAMPLE 3.3

Determining the Mode

A systems manager in charge of a company’s network keeps track of the number of server failures that occur in a day. Determine the mode for the following data, which represent the number of server failures per a day for the past two weeks:

1 3 0 3 26 2 7 4 0 2 3 3 6 3

(continued)

SOLUTION The ordered array for these data is

0 0 1 2 2 3 3 3 3 4 6 7 26

Because 3 occurs five times, more times than any other value, the mode is 3. Thus, the systems manager can say that the most common occurrence is having three server failures in a day. For this data set, the median is also equal to 3, and the mean is equal to 4.5. The value 26 is an extreme value. For these data, the median and the mode are better measures of central tendency than the mean.

The Geometric Mean

When you want to measure the rate of change of a variable over time, you need to use the geometric mean instead of the arithmetic mean. Equation (3.3) defines the geometric mean.

GEOMETRIC MEAN

The **geometric mean** is the n th root of the product of n values:

$$\bar{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n} \quad (3.3)$$

The geometric mean rate of return measures the average percentage return of an investment per time period. Equation (3.4) defines the geometric mean rate of return.

GEOMETRIC MEAN RATE OF RETURN

$$\bar{R}_G = [(1 + R_1) \times (1 + R_2) \times \cdots \times (1 + R_n)]^{1/n} - 1 \quad (3.4)$$

where

$$R_i = \text{rate of return in time period } i$$

To illustrate these measures, consider an investment of \$100,000 that declined to a value of \$50,000 at the end of Year 1 and then rebounded back to its original \$100,000 value at the end of Year 2. The rate of return for this investment per year for the two-year period is 0 because the starting and ending value of the investment is unchanged. However, the arithmetic mean of the yearly rates of return of this investment is

$$\bar{X} = \frac{(-0.50) + (1.00)}{2} = 0.25 \text{ or } 25\%$$

because the rate of return for Year 1 is

$$R_1 = \left(\frac{50,000 - 10,000}{100,000} \right) = -0.50 \text{ or } -50\%$$

and the rate of return for Year 2 is

$$R_2 = \left(\frac{10,000 - 50,000}{50,000} \right) = 1.00 \text{ or } 100\%$$

Using Equation (3.4), the geometric mean rate of return per year for the two years is

$$\begin{aligned}\bar{R}_G &= [(1 + R_1) \times (1 + R_2)]^{1/n} - 1 \\ &= [(1 + (-0.50)) \times (1 + (1.0))]^{1/2} - 1 \\ &= [(0.50) \times (2.0)]^{1/2} - 1 \\ &= [1.0]^{1/2} - 1 \\ &= 1 - 1 = 0\end{aligned}$$

Using the geometric mean rate of return more accurately reflects the (zero) change in the value of the investment per year for the two-year period than does the arithmetic mean.

EXAMPLE 3.4

Computing the Geometric Mean Rate of Return

The percentage change in the Russell 2000 Index of the stock prices of 2,000 small companies was -5.5% in 2011 and 14.6% in 2012. Compute the geometric rate of return.

SOLUTION Using Equation (3.4), the geometric mean rate of return in the Russell 2000 Index for the two years is

$$\begin{aligned}\bar{R}_G &= [(1 + R_1) \times (1 + R_2)]^{1/n} - 1 \\ &= [(1 + (-0.055)) \times (1 + (0.146))]^{1/2} - 1 \\ &= [(0.945) \times (1.146)]^{1/2} - 1 \\ &= (1.08297)^{1/2} - 1 \\ &= 1.0407 - 1 = 0.0407\end{aligned}$$

The geometric mean rate of return in the Russell 2000 Index for the two years is 4.07% per year.

3.2 Variation and Shape

In addition to central tendency, every variable can be characterized by its variation and shape. Variation measures the **spread**, or **dispersion**, of the values. One simple measure of variation is the range, the difference between the largest and smallest values. More commonly used in statistics are the standard deviation and variance, two measures explained later in this section. The shape of a variable represents a pattern of all the values, from the lowest to highest value. As you will learn later in this section, many variables have a pattern that looks approximately like a bell, with a peak of values somewhere in the middle.

The Range

The **range** is the difference between the largest and smallest value and is the simplest descriptive measure of variation for a numerical variable.

RANGE

The range is equal to the largest value minus the smallest value.

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}} \quad (3.5)$$

To further analyze the sample of 10 times to get ready in the morning, you can compute the range. To do so, you rank the data from smallest to largest:

29 31 35 39 39 40 43 44 44 52

Using Equation (3.5), the range is $52 - 29 = 23$ minutes. The range of 23 minutes indicates that the largest difference between any two days in the time to get ready in the morning is 23 minutes.

EXAMPLE 3.5

Computing the Range in the Calories in Cereals

Nutritional data about a sample of seven breakfast cereals (stored in **Cereals**) includes the number of calories per serving (see Example 3.1 on page 104). Compute the range of the number of calories for the cereals.

SOLUTION Ranked from smallest to largest, the calories for the seven cereals are

80 100 100 110 130 190 200

Therefore, using Equation (3.5), the range = $200 - 80 = 120$. The largest difference in the number of calories between any two cereals is 120.

The range measures the *total spread* in the set of data. Although the range is a simple measure of the total variation of the variable, it does not take into account *how* the values are distributed between the smallest and largest values. In other words, the range does not indicate whether the values are evenly distributed, clustered near the middle, or clustered near one or both extremes. Thus, using the range as a measure of variation when at least one value is an extreme value is misleading.

The Variance and the Standard Deviation

Being a simple measure of variation, the range does not consider how the values distribute or cluster between the extremes. Two commonly used measures of variation that account for how all the values are distributed are the **variance** and the **standard deviation**. These statistics measure the “average” scatter around the mean—how larger values fluctuate above it and how smaller values fluctuate below it.

A simple measure of variation around the mean might take the difference between each value and the mean and then sum these differences. However, if you did that, you would find that these differences sum to zero because the mean is the balance point for *every* numerical variable. A measure of variation that *differs* from one data set to another *squares* the difference between each value and the mean and then sums these squared differences. The sum of these squared differences, known as the **sum of squares (SS)**, is then used to compute the sample variance (S^2) and the sample standard deviation (S).

The **sample variance** (S^2) is the sum of squares divided by the sample size minus 1. The **sample standard deviation** (S) is the square root of the sample variance. Because this sum of squares will always be nonnegative according to the rules of algebra, *neither the variance nor the standard deviation can ever be negative*. For virtually all variables, the variance and standard deviation will be a positive value. Both of these statistics will be zero only if every value in the sample is the same value (i.e., the values show no variation).

For a sample containing n values, $X_1, X_2, X_3, \dots, X_n$, the sample variance (S^2) is

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n - 1}$$

Equations (3.6) and (3.7) define the sample variance and sample standard deviation using summation notation. The term $\sum_{i=1}^n (X_i - \bar{X})^2$ represents the sum of squares.

SAMPLE VARIANCE

The sample variance is the sum of the squared differences around the mean divided by the sample size minus 1:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \quad (3.6)$$

where

\bar{X} = sample mean

n = sample size

X_i = i th value of the variable X

$\sum_{i=1}^n (X_i - \bar{X})^2$ = summation of all the squared differences between the X_i values and \bar{X}

SAMPLE STANDARD DEVIATION

The sample standard deviation is the square root of the sum of the squared differences around the mean divided by the sample size minus 1:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}} \quad (3.7)$$

Note that in both equations, the sum of squares is divided by the sample size minus 1, $n - 1$. The value is used for reasons having to do with statistical inference and the properties of sampling distributions, a topic discussed in Section 7.2 on page 251. For now, observe that the difference between dividing by n and by $n - 1$ becomes smaller as the sample size increases.

In practice, you will most likely use the sample standard deviation as the measure of variation. Unlike the sample variance, a squared quantity, the standard deviation will always be a number expressed in the same units as the original sample data. For almost all sets of data, the majority of the values in a sample will be within an interval of plus and minus 1 standard deviation above and below the mean. Therefore, knowledge of the mean and the standard deviation usually helps define where at least the majority of the data values are clustering.

To hand-compute the sample variance, S^2 , and the sample standard deviation, S :

1. Compute the difference between each value and the mean.
2. Square each difference.
3. Sum the squared differences.
4. Divide this total by $n - 1$ to compute the sample variance.
5. Take the square root of the sample variance to compute the sample standard deviation.

To further analyze the sample of 10 times to get ready in the morning, Table 3.1 shows the first four steps for calculating the variance and standard deviation with a mean (\bar{X}) equal to 39.6. (Computing the mean is explained on page 103.) The second column of Table 3.1 shows step 1. The third column of Table 3.1 shows step 2. The sum of the squared differences (step 3) is shown at the bottom of Table 3.1. This total is then divided by $10 - 1 = 9$ to compute the variance (step 4).

Student Tip

Remember, neither the variance nor the standard deviation can ever be negative.

TABLE 3.1

Computing the Variance of the Getting-Ready Times

	Time (X)	<i>Step 1:</i> $(X_i - \bar{X})$	<i>Step 2:</i> $(X_i - \bar{X})^2$
	39	-0.60	0.36
	29	-10.60	112.36
	43	3.40	11.56
	52	12.40	153.76
$n = 10$	39	-0.60	0.36
$\bar{X} = 39.6$	44	4.40	19.36
	40	0.40	0.16
	31	-8.60	73.96
	44	4.40	19.36
	35	-4.60	21.16
	<i>Step 3: Sum</i>		412.40
	<i>Step 4: Divide by (n - 1)</i>		45.82

You can also compute the variance by substituting values for the terms in Equation (3.6):

$$\begin{aligned}
 S^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \\
 &= \frac{(39 - 39.6)^2 + (29 - 39.6)^2 + \dots + (35 - 39.6)^2}{10 - 1} \\
 &= \frac{412.4}{9} \\
 &= 45.82
 \end{aligned}$$

Because the variance is in squared units (in squared minutes, for these data), to compute the standard deviation, you take the square root of the variance. Using Equation (3.7) on page 109, the sample standard deviation, S , is

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}} = \sqrt{45.82} = 6.77$$

This indicates that the getting-ready times in this sample are clustering within 6.77 minutes around the mean of 39.6 minutes (i.e., clustering between $\bar{X} - 1S = 32.83$ and $\bar{X} + 1S = 46.37$). In fact, 7 out of 10 getting-ready times lie within this interval.

Using the second column of Table 3.1, you can also compute the sum of the differences between each value and the mean to be zero. For any set of data, this sum will always be zero:

$$\sum_{i=1}^n (X_i - \bar{X}) = 0 \text{ for all sets of data}$$

This property is one of the reasons that the mean is used as the most common measure of central tendency.

EXAMPLE 3.6
Computing the Variance and Standard Deviation of the Number of Calories in Cereals
TABLE 3.2

Computing the Variance of the Calories in the Cereals

Nutritional data about a sample of seven breakfast cereals (stored in **Cereals**) includes the number of calories per serving (see Example 3.1 on page 104). Compute the variance and standard deviation of the calories in the cereals.

SOLUTION Table 3.2 illustrates the computation of the variance and standard deviation for the calories in the cereals.

	Calories	Step 1: $(X_i - \bar{X})$	Step 2: $(X_i - \bar{X})^2$
$n = 7$ $\bar{X} = 130$	80	-50	2,500
	100	-30	900
	100	-30	900
	110	-20	400
	130	0	0
	190	60	3,600
	200	70	4,900
Step 3: Sum		13,200	
Step 4: Divide by $(n - 1)$		2,220	

Using Equation (3.6) on page 109:

$$\begin{aligned}
 S^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \\
 &= \frac{(80 - 130)^2 + (100 - 130)^2 + \dots + (200 - 130)^2}{7 - 1} \\
 &= \frac{13,200}{6} \\
 &= 2,200
 \end{aligned}$$

Using Equation (3.7) on page 109, the sample standard deviation, S , is

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}} = \sqrt{2,200} = 46.9042$$

The standard deviation of 46.9042 indicates that the calories in the cereals are clustering within ± 46.9042 around the mean of 130 (i.e., clustering between $\bar{X} - 1S = 83.0958$ and $\bar{X} + 1S = 176.9042$). In fact, 57.1% (four out of seven) of the calories lie within this interval.

The Coefficient of Variation

The coefficient of variation is equal to the standard deviation divided by the mean, multiplied by 100%. Unlike the measures of variation presented previously, the **coefficient of variation (CV)** measures the scatter in the data relative to the mean. The coefficient of variation is a *relative measure* of variation that is always expressed as a percentage rather than in terms of the units of the particular data. Equation (3.8) defines the coefficient of variation.

COEFFICIENT OF VARIATION

The coefficient of variation is equal to the standard deviation divided by the mean, multiplied by 100%.

Student Tip

The coefficient of variation is always expressed as a percentage, not in the units of the variables.

$$CV = \left(\frac{S}{\bar{X}} \right) 100\% \quad (3.8)$$

where

S = sample standard deviation
 \bar{X} = sample mean

For the sample of 10 getting-ready times, because $\bar{X} = 39.6$ and $S = 6.77$, the coefficient of variation is

$$CV = \left(\frac{S}{\bar{X}} \right) 100\% = \left(\frac{6.77}{39.6} \right) 100\% = 17.10\%$$

For the getting-ready times, the standard deviation is 17.1% of the size of the mean.

The coefficient of variation is especially useful when comparing two or more sets of data that are measured in different units, as Example 3.7 illustrates.

EXAMPLE 3.7

Comparing Two Coefficients of Variation When the Two Variables Have Different Units of Measurement

LEARN MORE

The Sharpe ratio, another relative measure of variation, is often used in financial analysis. Read the **SHORT TAKES** for Chapter 3 to learn more about this ratio.

Which varies more from cereal to cereal—the number of calories or the amount of sugar (in grams)?

SOLUTION Because calories and the amount of sugar have different units of measurement, you need to compare the relative variability in the two measurements.

For calories, using the mean and variance computed in Examples 3.1 and 3.6 on pages 104 and 111, the coefficient of variation is

$$CV_{\text{Calories}} = \left(\frac{46.9042}{130} \right) 100\% = 36.08\%$$

For the amount of sugar in grams, the values for the seven cereals are

6 2 4 4 4 11 10

For these data, $\bar{X} = 5.8571$ and $S = 3.3877$. Therefore, the coefficient of variation is

$$CV_{\text{Sugar}} = \left(\frac{3.3877}{5.8571} \right) 100\% = 57.84\%$$

You conclude that relative to the mean, the amount of sugar is much more variable than the calories.

Z Scores

The **Z score** of a value is the difference between that value and the mean, divided by the standard deviation. A Z score of 0 indicates that the value is the same as the mean. If a Z score is a positive or negative number, it indicates whether the value is above or below the mean and by how many standard deviations.

Z scores help identify **outliers**, the values that seem excessively different from most of the rest of the values (see Section 1.3). Values that are very different from the mean will have either very small (negative) Z scores or very large (positive) Z scores. As a general rule, a Z score that is less than -3.0 or greater than $+3.0$ indicates an outlier value.

Z SCORE

The Z score for a value is equal to the difference between the value and the mean, divided by the standard deviation:

$$Z = \frac{X - \bar{X}}{S} \quad (3.9)$$

To further analyze the sample of 10 times to get ready in the morning, you can compute the Z scores. Because the mean is 39.6 minutes, the standard deviation is 6.77 minutes, and the time to get ready on the first day is 39.0 minutes, you compute the Z score for Day 1 by using Equation (3.9):

$$\begin{aligned} Z &= \frac{X - \bar{X}}{S} \\ &= \frac{39.0 - 39.6}{6.77} \\ &= -0.09 \end{aligned}$$

The Z score of -0.09 for the first day indicates that the time to get ready on that day is very close to the mean. Table 3.3 presents the Z scores for all 10 days.

TABLE 3.3

Z Scores for the 10 Getting-Ready Times

	Time (X)	Z Score
	39	-0.09
	29	-1.57
	43	0.50
	52	1.83
$\bar{X} = 39.6$	39	-0.09
$S = 6.77$	44	0.65
	40	0.06
	31	-1.27
	44	0.65
	35	-0.68

The largest Z score is 1.83 for Day 4, on which the time to get ready was 52 minutes. The lowest Z score is -1.57 for Day 2, on which the time to get ready was 29 minutes. Because none of the Z scores are less than -3.0 or greater than $+3.0$, you conclude that the getting-ready times include no apparent outliers.

EXAMPLE 3.8**Computing the Z Scores of the Number of Calories in Cereals**

Nutritional data about a sample of seven breakfast cereals (stored in **Cereals**) includes the number of calories per serving (see Example 3.1 on page 104). Compute the Z scores of the calories in breakfast cereals.

SOLUTION Table 3.4 presents the Z scores of the calories for the cereals. The largest Z score is 1.49, for a cereal with 200 calories. The lowest Z score is -1.07 , for a cereal with 80 calories. There are no apparent outliers in these data because none of the Z scores are less than -3.0 or greater than $+3.0$.

TABLE 3.4

Z Scores of the Number of Calories in Cereals

	Calories	Z Scores
	80	-1.07
	100	-0.64
$\bar{X} = 130$	100	-0.64
$S = 46.9042$	110	-0.43
	130	0.00
	190	1.28
	200	1.49

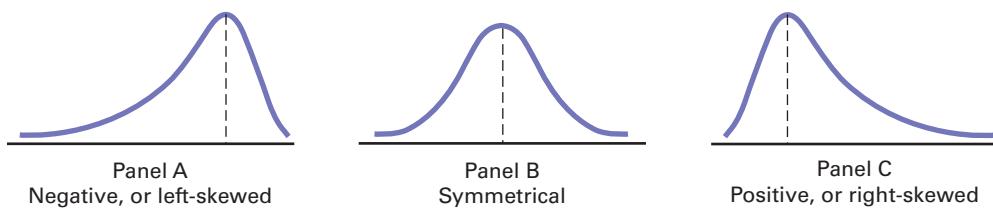
Shape: Skewness and Kurtosis

The pattern to the distribution of values throughout the entire range of all the values is called the **shape**. The shape of the distribution of data values can be described by two statistics: skewness and kurtosis.

Skewness measures the extent to which the data values are not **symmetrical** around the mean. In a perfectly symmetrical distribution, the values below the mean are distributed in exactly the same way as the values above the mean, and the skewness is zero. In a **skewed** distribution, there is an imbalance of data values below and above the mean, and the skewness is a nonzero value. Figure 3.1 depicts the shape of the distribution of values for three variables, with the mean for each variable plotted as a dashed vertical line.

FIGURE 3.1

The shapes of three data distributions



In Panel A, the distribution of values is **left-skewed**. In this panel, most of the values are in the upper portion of the distribution. A long tail and distortion to the left is caused by some extremely small values. Because the skewness statistic for such a distribution will be less than zero, the term *negative skew* is also used to describe this distribution. These extremely small values pull the mean downward so that the mean is less than the median.

In Panel B, the distribution of values is **symmetrical**. The portion of the curve below the mean is the mirror image of the portion of the curve above the mean. There is no asymmetry of data values below and above the mean, the mean equals the median, and, as noted earlier, the skewness is zero.

In Panel C, the distribution of values is **right-skewed**. In this panel, most of the values are in the lower portion of the distribution. A long tail on the right is caused by some extremely large values. Because the skewness statistic for such a distribution will be greater than zero, the term *positive skew* is also used to describe this distribution. These extremely large values pull the mean upward so that the mean is greater than the median.

The observations about the mean and median made when examining Figure 3.1 generally hold for most distributions of a continuous numerical variable. Summarized, these observations are:

- **Mean < median:** negative, or left-skewed distribution
- **Mean = median:** symmetrical distribution with zero skewness
- **Mean > median:** positive, or right-skewed distribution

Kurtosis measures the extent to which values that are very different from the mean affect the shape of the distribution of a set of data. Kurtosis affects the peakedness of the curve of the distribution—that is, how sharply the curve rises approaching the center of the distribution. Kurtosis compares the shape of the peak to the shape of the peak of a normal distribution (discussed in Chapter 6), which, by definition, has a kurtosis of zero.¹ A distribution that has a sharper-rising center peak than the peak of a normal distribution has *positive* kurtosis, a kurtosis value that is greater than zero, and is called **lepkurtic**. A distribution that has a slower-rising (flatter) center peak than the peak of a normal distribution has *negative* kurtosis, a kurtosis value that is less than zero, and is called **platykurtic**. A lepkurtic distribution has a higher concentration of values near the mean of the distribution compared to a normal distribution, while a platykurtic distribution has a lower concentration compared to a normal distribution.

In affecting the shape of the central peak, the relative concentration of values near the mean also affects the ends, or *tails*, of the curve of a distribution. A lepkurtic distribution has *fatter* tails, many more values in the tails, than a normal distribution has. If decision making about a set of data mistakenly assumes a normal distribution, when, in fact, the data forms a lepkurtic distribution, then that decision making will underestimate the occurrence of extreme values (values that are very different from the mean). Such an observation has been a basis for several explanations about the unanticipated reverses and collapses that financial markets have experienced in the recent past. (See reference 5 for an example of such an explanation.)

¹Several different operational definitions exist for kurtosis. The definition here, used by both Excel and Minitab, is sometimes called *excess kurtosis* to distinguish it from other definitions. Read the SHORT TAKES for Chapter 3 to learn how Excel calculates kurtosis (and skewness).

EXAMPLE 3.9

Descriptive Statistics for Growth and Value Funds

In the More Descriptive Choices scenario, you are interested in comparing the past performance of the growth and value funds from a sample of 316 funds. One measure of past performance is the one-year return percentage variable. Compute descriptive statistics for the growth and value funds.

SOLUTION Figure 3.2 presents descriptive summary measures for the two types of funds. The results include the mean, median, mode, minimum, maximum, range, variance, standard deviation, coefficient of variation, skewness, kurtosis, count (the sample size), and standard error. The standard error, discussed in Section 7.2, is the standard deviation divided by the square root of the sample size.

FIGURE 3.2

Excel and Minitab Descriptive statistics for the one-year return percentages for the growth and value funds

A		B		C		Descriptive Statistics: 1YrReturn%								
1	Descriptive Statistics for the 1YrReturn% Variable													
2														
3														
4	Growth													
5	Value													
6	Mean													
7	14.28													
8	Median													
9	14.18													
10	Mode													
11	16.95													
12	Minimum													
13	-11.28													
14	Maximum													
15	33.98													
16	Range													
17	45.26													
18	26.6													
19	Variance													
20	25.0413													
21	19.9369													
22	Standard Deviation													
23	5.0041													
24	4.4651													
25	Coeff. of Variation													
26	35.05%													
27	30.38%													
28	Skewness													
29	0.2039													
30	Kurtosis													
31	5.1479													
32	0.6684													
33	Count													
34	227													
35	89													
36	Standard Error													
37	0.3321													
38	0.4733													

In examining the results, you see that there are some differences in the one-year return for the growth and value funds. The growth funds had a mean one-year return of 14.28 and a median return of 14.18. This compares to a mean of 14.70 and a median of 15.30 for the value funds.

(continued)

funds. The medians indicate that half of the growth funds had one-year returns of 14.18 or better, and half the value funds had one-year returns of 15.30 or better. You conclude that the value funds had a slightly higher return than the growth funds.

The growth funds had a higher standard deviation than the value funds (5.0041, as compared to 4.4651). Both the growth funds and the value funds showed very little skewness, as the skewness of the growth funds was 0.2039 and the skewness of the value funds was -0.2083. The kurtosis of the growth funds was very positive, indicating a distribution that was much more peaked than a normal distribution. The kurtosis of the value funds was slightly positive indicating a distribution that did not depart markedly from a normal distribution.

EXAMPLE 3.10

Descriptive Statistics Using Multidimensional Contingency Tables

Continuing with the More Descriptive Choices scenario, you wish to explore the effects of each combination of type, market cap, and risk on measures of past performance. One measure of past performance is the three-year return percentage. Compute the mean three-year return percentage for each combination of type, market cap, and risk.

SOLUTION Compute the mean for each combination by adding the numerical variable three-year return percentage to a multidimensional contingency table, first introduced in Section 2.6 on page 69 as a way of summarizing many categorical variables. The **Add Numerical Variable to MCT online topic** discusses this technique, using this example to explain how to use Excel or Minitab instructions to construct such a table. Shown below are the Excel and Minitab tables for this example.

Type	Low	Average	High	Grand Total
Growth	10.65	10.20	11.88	10.56
Large	9.56	8.83	22.17	9.86
Mid-Cap	12.10	10.28	*	11.21
Small	13.41	10.58	7.47	11.03
Value	9.49	10.24	9.30	9.63
Large	8.76	6.94	6.70	8.57
Mid-Cap	10.56	11.65	*	10.79
Small	11.33	11.08	10.61	11.15
Grand Total	10.27	10.21	11.29	10.29

Tabulated statistics: Type, Market Cap, Risk

Rows: Type / Market Cap Columns: Risk

	Average	High	Low	All
Growth				
Large	8.83	22.17	9.56	9.86
Mid-Cap	10.28	*	12.10	11.21
Small	10.58	7.47	13.41	11.03
Value				
Large	6.94	6.70	8.76	8.57
Mid-Cap	11.65	*	10.56	10.79
Small	11.08	10.61	11.33	11.15
All	10.21	11.29	10.27	10.29
	91	13	212	316

Cell Contents: 3YrReturn% : Mean
Count

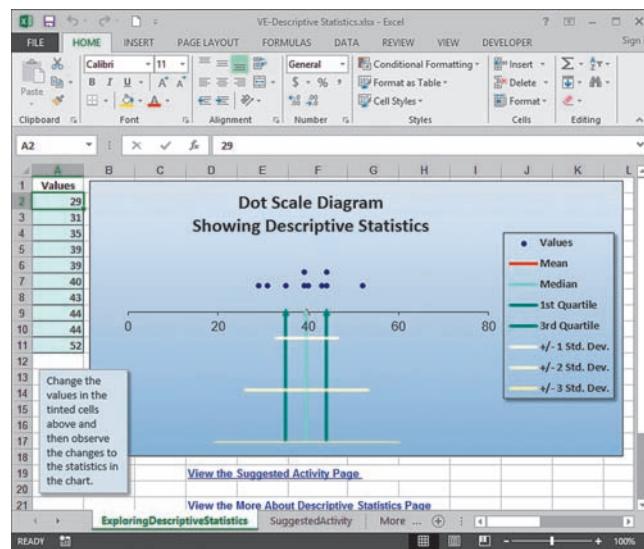
Analyzing each combination of type, market cap, and risk reveals patterns that would not be seen if the mean of the three-year return percentage had been computed for only the growth and value funds (similar to what is done in Example 3.9). Empty cells (Excel) and starred cells (Minitab), such as those for mid-cap growth funds with high risk, represent combinations that do not exist in the sample of 316 funds.

VISUAL EXPLORATIONS

Exploring Descriptive Statistics

Open the **VE-Descriptive Statistics workbook** to explore the effects of changing data values on measures of central tendency, variation, and shape. Change the data values in the cell range **A2:A11** and then observe the changes to the statistics shown in the chart.

Click **View the Suggested Activity Page** to view a specific change you could make to the data values in column A. Click **View the More About Descriptive Statistics Page** to view summary definitions of the descriptive statistics shown in the chart. (See Appendix C to learn how you can download a copy of this workbook.)



Problems for Sections 3.1 and 3.2

LEARNING THE BASICS

3.1 The following set of data is from a sample of $n = 5$:

$$7 \ 4 \ 9 \ 8 \ 2$$

- Compute the mean, median, and mode.
- Compute the range, variance, standard deviation, and coefficient of variation.
- Compute the Z scores. Are there any outliers?
- Describe the shape of the data set.

3.2 The following set of data is from a sample of $n = 6$:

$$7 \ 4 \ 9 \ 7 \ 3 \ 12$$

- Compute the mean, median, and mode.
- Compute the range, variance, standard deviation, and coefficient of variation.
- Compute the Z scores. Are there any outliers?
- Describe the shape of the data set.

3.3 The following set of data is from a sample of $n = 7$:

$$12 \ 7 \ 4 \ 9 \ 0 \ 7 \ 3$$

- Compute the mean, median, and mode.
- Compute the range, variance, standard deviation, and coefficient of variation.
- Compute the Z scores. Are there any outliers?
- Describe the shape of the data set.

3.4 The following set of data is from a sample of $n = 5$:

$$7 \ -5 \ -8 \ 7 \ 9$$

- Compute the mean, median, and mode.

- Compute the range, variance, standard deviation, and coefficient of variation.
- Compute the Z scores. Are there any outliers?
- Describe the shape of the data set.

3.5 Suppose that the rate of return for a particular stock during the past two years was 10% for one of the years and 30% for the other year. Compute the geometric rate of return per year. (*Note:* A rate of return of 10% is recorded as 0.10, and a rate of return of 30% is recorded as 0.30.)

3.6 Suppose that the rate of return for a particular stock during the past two years was 20% for one of the years and -30% for the other year. Compute the geometric rate of return per year.

APPLYING THE CONCEPTS

3.7 *Saveur*, a gourmet food, wine, and travel magazine, reported the following summary for the household incomes of its two types of subscribers—the print reader and the digital reader.

Audience	Median
Saveur reader	\$163,108
Saveur.com reader	84,548

Source: Data extracted from *Saveur* 2013 Advertising Media Kit, bit.ly/14tiv4P.

Interpret the median household income for the *Saveur* readers and the **Saveur.com** readers.

3.8 The operations manager of a plant that manufactures tires wants to compare the actual inner diameters of two grades of tires, each of which is expected to be 575 millimeters. A sample of five tires of each grade was selected, and the results representing the inner diameters of the tires, ranked from smallest to largest, are as follows:

Grade X	Grade Y
568 570 575 578 584	573 574 575 577 578

- a. For each of the two grades of tires, compute the mean, median, and standard deviation.
- b. Which grade of tire is providing better quality? Explain.
- c. What would be the effect on your answers in (a) and (b) if the last value for grade Y was 588 instead of 578? Explain.

3.9 According to the U.S. Census Bureau, in 2012, the median sales price of new houses was \$245,200, and the mean sales price was \$291,200 (extracted from www.census.gov, March 14, 2013).

- a. Interpret the median sales price.
- b. Interpret the mean sales price.
- c. Discuss the shape of the distribution of the price of new houses.

 **3.10** The file **FastFood** contains the amount that a sample of 15 customers spent for lunch (\$) at a fast-food restaurant:

7.42	6.29	5.83	6.50	8.34	9.51	7.10	6.80
5.90	4.89	6.50	5.52	7.90	8.30	9.60	

- a. Compute the mean and median.
- b. Compute the variance, standard deviation, range, and coefficient of variation.
- c. Are the data skewed? If so, how?
- d. Based on the results of (a) through (c), what conclusions can you reach concerning the amount that customers spent for lunch?

3.11 The file **Sedans** contains the overall miles per gallon (MPG) of 2013 midsized sedans:

38	26	30	26	25	27	22	27	39	24	24	26	25
23	25	26	31	26	37	22	29	25	33	21	21	

Source: Data extracted from "Ratings," *Consumer Reports*, April 2013, pp. 30–31.

- a. Compute the mean, median, and mode.
- b. Compute the variance, standard deviation, range, coefficient of variation, and Z scores.
- c. Are the data skewed? If so, how?
- d. Compare the results of (a) through (c) to those of Problem 3.12 (a) through (c) that refer to the miles per gallon of small SUVs.

3.12 The file **SUV** contains the overall miles per gallon (MPG) of 2013 small SUVs:

22	23	21	22	25	26	22	22	21
19	22	22	26	23	24	21	22	

Source: Data extracted from "Ratings," *Consumer Reports*, April 2013, pp. 34–35.

- a. Compute the mean, median, and mode.
- b. Compute the variance, standard deviation, range, coefficient of variation, and Z scores.
- c. Are the data skewed? If so, how?
- d. Compare the results of (a) through (c) to those of Problem 3.11 (a) through (c) that refer to the MPG of midsized sedans.

3.13 The file **AccountingPartners** contains the number of partners in a cohort of rising accounting firms that have been tagged as "firms to watch." The firms have the following numbers of partners:

25	18	23	16	16	14	22	17	9	32	22	12	18	9	25	9	28	14
22	18	22	16	12	31	17	12	14	16	30	12	12	10	24	12	11	

Source: Data extracted from bit.ly/11asPHm.

- a. Compute the mean, median, and mode.
- b. Compute the variance, standard deviation, range, coefficient of variation, and Z scores. Are there any outliers? Explain.
- c. Are the data skewed? If so, how?
- d. Based on the results of (a) through (c), what conclusions can you reach concerning the number of partners in rising accounting firms?

3.14 The file **MarketPenetration** contains the market penetration value (the percentage of the country population that are users) for the 15 countries that lead the world in total number of Facebook users:

52.56	33.09	5.37	19.41	32.52	41.69	51.61	30.12
39.07	30.62	38.16	49.35	27.13	53.45	40.01	

Source: Data extracted from www.socialbakers.com/facebook-statistics/.

- a. Compute the mean, median, and mode.
- b. Compute the variance, standard deviation, range, coefficient of variation, and Z scores. Are there any outliers? Explain.
- c. Are the data skewed? If so, how?
- d. Based on the results of (a) through (c), what conclusions can you reach concerning Facebook's market penetration?

3.15 Is there a difference in the variation of the yields of different types of investments? The file **CD Rate** contains the yields for one-year certificates of deposit (CDs) and five-year CDs for 23 banks in the United States, as of March 20, 2013.

Source: Data extracted from www.Bankrate.com, March 20, 2013.

- a. For one-year and five-year CDs, separately compute the variance, standard deviation, range, and coefficient of variation.
- b. Based on the results of (a), do one-year CDs or five-year CDs have more variation in the yields offered? Explain.

3.16 The file **HotelAway** contains the average room price (in English pounds) paid in 2012 by people of various nationalities while traveling away from their home country:

171	169	158	164	158	135	150	141
-----	-----	-----	-----	-----	-----	-----	-----

Source: Data extracted from bit.ly/XPtF0Y.

- a. Compute the mean, median, and mode.
- b. Compute the range, variance, and standard deviation.
- c. Based on the results of (a) and (b), what conclusions can you reach concerning the room price (in US\$) in 2012?
- d. Suppose that the first value was 200 instead of 171. Repeat (a) through (c), using this value. Comment on the difference in the results.

3.17 A bank branch located in a commercial district of a city has the business objective of developing an improved process for serving customers during the noon-to-1:00 P.M. lunch period. The waiting time, in minutes, is defined as the time the customer

enters the line to when he or she reaches the teller window. Data collected from a sample of 15 customers during this hour are stored in **Bank1**:

4.21	5.55	3.02	5.13	4.77	2.34	3.54	3.20
4.50	6.10	0.38	5.12	6.46	6.19	3.79	

- Compute the mean and median.
- Compute the variance, standard deviation, range, coefficient of variation, and Z scores. Are there any outliers? Explain.
- Are the data skewed? If so, how?
- As a customer walks into the branch office during the lunch hour, she asks the branch manager how long she can expect to wait. The branch manager replies, "Almost certainly less than five minutes." On the basis of the results of (a) through (c), evaluate the accuracy of this statement.

3.18 Suppose that another bank branch, located in a residential area, is also concerned with the noon-to-1:00 P.M. lunch hour. The waiting time, in minutes, collected from a sample of 15 customers during this hour, are stored in **Bank2**:

9.66	5.90	8.02	5.79	8.73	3.82	8.01	8.35
10.49	6.68	5.64	4.08	6.17	9.91	5.47	

- Compute the mean and median.
- Compute the variance, standard deviation, range, coefficient of variation, and Z scores. Are there any outliers? Explain.
- Are the data skewed? If so, how?
- As a customer walks into the branch office during the lunch hour, he asks the branch manager how long he can expect to wait. The branch manager replies, "Almost certainly less than five minutes." On the basis of the results of (a) through (c), evaluate the accuracy of this statement.

3.19 General Electric (GE) is one of the world's largest companies; it develops, manufactures, and markets a wide range of products, including medical diagnostic imaging devices, jet engines, lighting products, and chemicals. In 2011, the stock price rose 1.4%, and in 2012, the stock price rose 17.2%.

Source: Data extracted from [finance.yahoo.com](#), June 24, 2012.

- Compute the geometric mean rate of return per year for the two-year period 2011–2012. (Hint: Denote an increase of 1.4% as $R_2 = 0.014$.)
- If you purchased \$1,000 of GE stock at the start of 2011, what was its value at the end of 2012?
- Compare the result of (b) to that of Problem 3.20 (b).

 **3.20** TASER International, Inc., develops, manufactures, and sells nonlethal self-defense devices known as Tasers and markets primarily to law enforcement, corrections institutions, and the military. TASER's stock price in 2011 increased by 2.4%, and in 2012, it increased by 74.6%.

Source: Data extracted from [finance.yahoo.com](#), March 29, 2013.

- Compute the geometric mean rate of return per year for the two-year period 2011–2012. (Hint: Denote an increase of 2.4% as $R_1 = 0.024$.)
- If you purchased \$1,000 of TASER stock at the start of 2011, what was its value at the end of 2012?
- Compare the result of (b) to that of Problem 3.19 (b).

3.21 The file **Indices** contains data that represent the yearly rate of return (in percentage) for the Dow Jones Industrial

Average (DJIA), the Standard & Poor's 500 (S&P 500), and the technology-heavy NASDAQ Composite (NASDAQ) from 2009 through 2012. These data are:

Year	DJIA	S&P 500	NASDAQ
2012	7.26	13.41	15.91
2011	5.50	0.00	-1.80
2010	11.00	12.80	16.90
2009	18.80	23.50	43.90

Source: Data extracted from [finance.yahoo.com](#), March 29, 2013.

- Compute the geometric mean rate of return per year for the DJIA, S&P 500, and NASDAQ from 2009 through 2012.
- What conclusions can you reach concerning the geometric mean rates of return per year of the three market indices?
- Compare the results of (b) to those of Problem 3.22 (b).

3.22 In 2009 through 2012, the value of precious metals fluctuated dramatically. The data in the following table (contained in the file **Metals**) represent the yearly rate of return (in percentage) for platinum, gold, and silver from 2009 through 2012:

Year	Platinum	Gold	Silver
2012	8.7	0.1	7.1
2011	-21.1	10.2	-9.8
2010	21.5	29.8	83.7
2009	55.9	23.9	49.3

Source: Data extracted from [finance.yahoo.com](#), March 29, 2013.

- Compute the geometric mean rate of return per year for platinum, gold, and silver from 2009 through 2012.
- What conclusions can you reach concerning the geometric mean rates of return of the three precious metals?
- Compare the results of (b) to those of Problem 3.21 (b).

3.23 Using the one-year return percentage variable in **Retirement Funds**:

- Construct a table that computes the mean for each combination of type, market cap, and risk.
- Construct a table that computes the standard deviation for each combination of type, market cap, and risk.
- What conclusions can you reach concerning differences among the types of retirement funds (growth and value), based on market cap (small, mid-cap, and large) and the risk (low, average, and high)?

3.24 Using the one-year return percentage variable in **Retirement Funds**:

- Construct a table that computes the mean for each combination of type, market cap, and rating.
- Construct a table that computes the standard deviation for each combination of type, market cap, and rating.
- What conclusions can you reach concerning differences among the types of retirement funds (growth and value), based on market cap (small, mid-cap, and large) and the rating (one, two, three, four, and five)?

3.25 Using the one-year return percentage variable in **Retirement Funds**:

- Construct a table that computes the mean for each combination of market cap, risk, and rating.
- Construct a table that computes the standard deviation for each combination of market cap, risk, and rating.
- What conclusions can you reach concerning differences based on the market cap (small, mid-cap, and large), risk (low, average, and high), and rating (one, two, three, four, and five)?

3.26 Using the one-year return percentage variable in **Retirement Funds**:

- Construct a table that computes the mean for each combination of type, risk, and rating.
- Construct a table that computes the standard deviation for each combination of type, risk, and rating.
- What conclusions can you reach concerning differences among the types of retirement funds (growth and value), based on the risk (low, average, and high) and the rating (one, two, three, four, and five)?

3.3 Exploring Numerical Data

Sections 3.1 and 3.2 discuss measures of central tendency, variation, and shape. You can also visualize the distribution of the values for a numerical variable by computing the quartiles and the five-number summary and constructing a boxplot.

Quartiles

Quartiles split the values into four equal parts—the **first quartile (Q_1)** divides the smallest 25.0% of the values from the other 75.0% that are larger. The **second quartile (Q_2)** is the median; 50.0% of the values are smaller than or equal to the median, and 50.0% are larger than or equal to the median. The **third quartile (Q_3)** divides the smallest 75.0% of the values from the largest 25.0%. Equations (3.10) and (3.11) define the first and third quartiles.

Student Tip

The methods of this section are commonly used in **exploratory data analysis**.

FIRST QUARTILE, Q_1

25.0% of the values are smaller than or equal to Q_1 , the first quartile, and 75.0% are larger than or equal to the first quartile, Q_1 :

$$Q_1 = \frac{n + 1}{4} \text{ ranked value} \quad (3.10)$$

THIRD QUARTILE, Q_3

75.0% of the values are smaller than or equal to the third quartile, Q_3 , and 25.0% are larger than or equal to the third quartile, Q_3 :

$$Q_3 = \frac{3(n + 1)}{4} \text{ ranked value} \quad (3.11)$$

Student Tip

As is the case when you compute the median, you must rank the values in order from smallest to largest before computing the quartiles.

Use the following rules to compute the quartiles from a set of ranked values:

- **Rule 1** If the ranked value is a whole number, the quartile is equal to the measurement that corresponds to that ranked value. For example, if the sample size $n = 7$, the first quartile, Q_1 , is equal to the measurement associated with the $(7 + 1)/4 =$ second ranked value.

- **Rule 2** If the ranked value is a fractional half (2.5, 4.5, etc.), the quartile is equal to the measurement that corresponds to the average of the measurements corresponding to the two ranked values involved. For example, if the sample size $n = 9$, the first quartile, Q_1 , is equal to the $(9 + 1)/4 = 2.5$ ranked value, halfway between the second ranked value and the third ranked value.
- **Rule 3** If the ranked value is neither a whole number nor a fractional half, you round the result to the nearest integer and select the measurement corresponding to that ranked value. For example, if the sample size $n = 10$, the first quartile, Q_1 , is equal to the $(10 + 1)/4 = 2.75$ ranked value. Round 2.75 to 3 and use the third ranked value.

To further analyze the sample of 10 times to get ready in the morning, you can compute the quartiles. To do so, you rank the data from smallest to largest:

<i>Ranked values:</i>	29	31	35	39	39	40	43	44	44	52
<i>Ranks:</i>	1	2	3	4	5	6	7	8	9	10

The first quartile is the $(n + 1)/4 = (10 + 1)/4 = 2.75$ ranked value. Using Rule 3, you round up to the third ranked value. The third ranked value for the getting-ready data is 35 minutes. You interpret the first quartile of 35 to mean that on 25% of the days, the time to get ready is less than or equal to 35 minutes, and on 75% of the days, the time to get ready is greater than or equal to 35 minutes.

The third quartile is the $3(n + 1)/4 = 3(10 + 1)/4 = 8.25$ ranked value. Using Rule 3 for quartiles, you round this down to the eighth ranked value. The eighth ranked value is 44 minutes. Thus, on 75% of the days, the time to get ready is less than or equal to 44 minutes, and on 25% of the days, the time to get ready is greater than or equal to 44 minutes.

Percentiles Related to quartiles are **percentiles** that split a variable into 100 equal parts. By this definition, the first quartile is equivalent to the 25th percentile, the second quartile to the 50th percentile, and the third quartile to the 75th percentile. Learn more about percentiles in the SHORT TAKES for Chapter 3.

EXAMPLE 3.11

Computing the Quartiles

Nutritional data about a sample of seven breakfast cereals (stored in **Cereals**) includes the number of calories per serving (see Example 3.1 on page 104). Compute the first quartile (Q_1) and third quartile (Q_3) of the number of calories for the cereals.

SOLUTION Ranked from smallest to largest, the numbers of calories for the seven cereals are as follows:

<i>Ranked values:</i>	80	100	100	110	130	190	200
<i>Ranks:</i>	1	2	3	4	5	6	7

For these data

$$\begin{aligned} Q_1 &= \frac{(n + 1)}{4} \text{ ranked value} \\ &= \frac{7 + 1}{4} \text{ ranked value} = 2\text{nd ranked value} \end{aligned}$$

Therefore, using Rule 1, Q_1 is the second ranked value. Because the second ranked value is 100, the first quartile, Q_1 , is 100.

(continued)

To compute the third quartile, Q_3 ,

$$\begin{aligned} Q_3 &= \frac{3(n + 1)}{4} \text{ ranked value} \\ &= \frac{3(7 + 1)}{4} \text{ ranked value} = 6\text{th ranked value} \end{aligned}$$

Therefore, using Rule 1, Q_3 is the sixth ranked value. Because the sixth ranked value is 190, Q_3 is 190.

The first quartile of 100 indicates that 25% of the cereals contain 100 calories or fewer per serving and 75% contain 100 or more calories. The third quartile of 190 indicates that 75% of the cereals contain 190 calories or fewer per serving and 25% contain 190 or more calories.

The Interquartile Range

The **interquartile range** (also called the **midspread**) measures the difference in the center of a distribution between the third and first quartiles.

INTERQUARTILE RANGE

The interquartile range is the difference between the third quartile and the first quartile:

$$\text{Interquartile range} = Q_3 - Q_1 \quad (3.12)$$

The interquartile range measures the spread in the middle 50% of the values. Therefore, it is not influenced by extreme values. To further analyze the sample of 10 times to get ready in the morning, you can compute the interquartile range. You first order the data as follows:

29 31 35 39 39 40 43 44 44 52

You use Equation (3.12) and the earlier results on page 121, $Q_1 = 35$ and $Q_3 = 44$:

$$\text{Interquartile range} = 44 - 35 = 9 \text{ minutes}$$

Therefore, the interquartile range in the time to get ready is 9 minutes. The interval 35 to 44 is often referred to as the *middle fifty*.

EXAMPLE 3.12

Computing the Interquartile Range for the Number of Calories in Cereals

Nutritional data about a sample of seven breakfast cereals (stored in **Cereals**) includes the number of calories per serving (see Example 3.1 on page 104). Compute the interquartile range of the number of calories in cereals.

SOLUTION Ranked from smallest to largest, the numbers of calories for the seven cereals are as follows:

80 100 100 110 130 190 200

Using Equation (3.12) and the earlier results from Example 3.11 on page 121, $Q_1 = 100$ and $Q_3 = 190$:

$$\text{Interquartile range} = 190 - 100 = 90$$

Therefore, the interquartile range of the number of calories in cereals is 90 calories.

Because the interquartile range does not consider any value smaller than Q_1 or larger than Q_3 , it cannot be affected by extreme values. Descriptive statistics such as the median, Q_1 , Q_3 , and the interquartile range, which are not influenced by extreme values, are called **resistant measures**.

The Five-Number Summary

The **five-number summary** for a variable consists of the smallest value (X_{smallest}), the first quartile, the median, the third quartile, and the largest value (X_{largest}).

FIVE-NUMBER SUMMARY

X_{smallest} Q_1 Median Q_3 X_{largest}

The five-number summary provides a way to determine the shape of the distribution for a set of data. Table 3.5 explains how relationships among these five statistics help to identify the shape of the distribution.

TABLE 3.5

Relationships Among the Five-Number Summary and the Type of Distribution

COMPARISON	TYPE OF DISTRIBUTION		
	Left-Skewed	Symmetrical	Right-Skewed
The distance from X_{smallest} to the median versus the distance from the median to X_{largest} .	The distance from X_{smallest} to the median is greater than the distance from the median to X_{largest} .	The two distances are the same.	The distance from X_{smallest} to the median is less than the distance from the median to X_{largest} .
The distance from X_{smallest} to Q_1 versus the distance from Q_3 to X_{largest} .	The distance from X_{smallest} to Q_1 is greater than the distance from Q_3 to X_{largest} .	The two distances are the same.	The distance from X_{smallest} to Q_1 is less than the distance from Q_3 to X_{largest} .
The distance from Q_1 to the median versus the distance from the median to Q_3 .	The distance from Q_1 to the median is greater than the distance from the median to Q_3 .	The two distances are the same.	The distance from Q_1 to the median is less than the distance from the median to Q_3 .

To further analyze the sample of 10 times to get ready in the morning, you can compute the five-number summary. For these data, the smallest value is 29 minutes, and the largest value is 52 minutes (see page 105). Calculations done on pages 105 and 121 show that the median = 39.5, Q_1 = 35, and Q_3 = 44. Therefore, the five-number summary is as follows:

29 35 39.5 44 52

The distance from X_{smallest} to the median ($39.5 - 29 = 10.5$) is slightly less than the distance from the median to X_{largest} ($52 - 39.5 = 12.5$). The distance from X_{smallest} to Q_1 ($35 - 29 = 6$) is slightly less than the distance from Q_3 to X_{largest} ($52 - 44 = 8$). The distance from Q_1 to the median ($39.5 - 35 = 4.5$) is the same as the distance from the median to Q_3 ($44 - 39.5 = 4.5$). Therefore, the getting-ready times are slightly right-skewed.

EXAMPLE 3.13

Computing the Five-Number Summary of the Number of Calories in Cereals in Cereals

Nutritional data about a sample of seven breakfast cereals (stored in **Cereals**) includes the number of calories per serving (see Example 3.1 on page 104). Compute the five-number summary of the number of calories in cereals.

SOLUTION From previous computations for the number of calories in cereals (see pages 105 and 121–122), you know that the median = 110, Q_1 = 100, and Q_3 = 190.

(continued)

In addition, the smallest value in the data set is 80, and the largest value is 200. Therefore, the five-number summary is as follows:

80 100 110 190 200

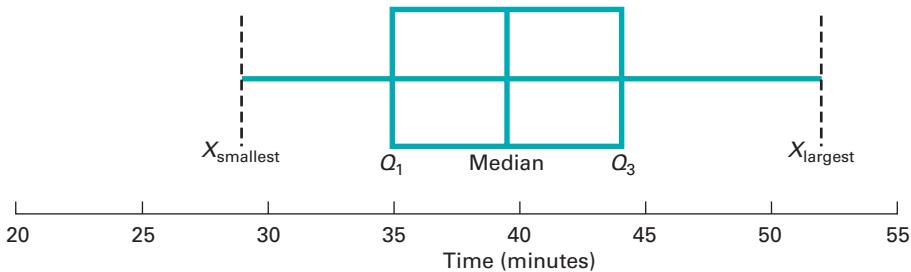
The three comparisons listed in Table 3.5 are used to evaluate skewness. The distance from X_{smallest} to the median ($110 - 80 = 30$) is less than the distance ($200 - 110 = 90$) from the median to X_{largest} . The distance from X_{smallest} to Q_1 ($100 - 80 = 20$) is greater than the distance from Q_3 to X_{largest} ($200 - 190 = 10$). The distance from Q_1 to the median ($110 - 100 = 10$) is less than the distance from the median to Q_3 ($190 - 110 = 80$). Two comparisons indicate a right-skewed distribution, whereas the other indicates a left-skewed distribution. Therefore, given the small sample size and the conflicting results, the shape is not clearly determined.

The Boxplot

The **boxplot** uses a five-number summary to visualize the shape of the distribution for a variable. Figure 3.3 contains a boxplot for the sample of 10 times to get ready in the morning.

FIGURE 3.3

Boxplot for the getting-ready times



The vertical line drawn within the box represents the median. The vertical line at the left side of the box represents the location of Q_1 , and the vertical line at the right side of the box represents the location of Q_3 . Thus, the box contains the middle 50% of the values. The lower 25% of the data are represented by a line connecting the left side of the box to the location of the smallest value, X_{smallest} . Similarly, the upper 25% of the data are represented by a line connecting the right side of the box to X_{largest} .

The Figure 3.3 boxplot for the getting-ready times shows a slight right-skewness: The distance between the median and the highest value is slightly greater than the distance between the lowest value and the median, and the right tail is slightly longer than the left tail.

EXAMPLE 3.14

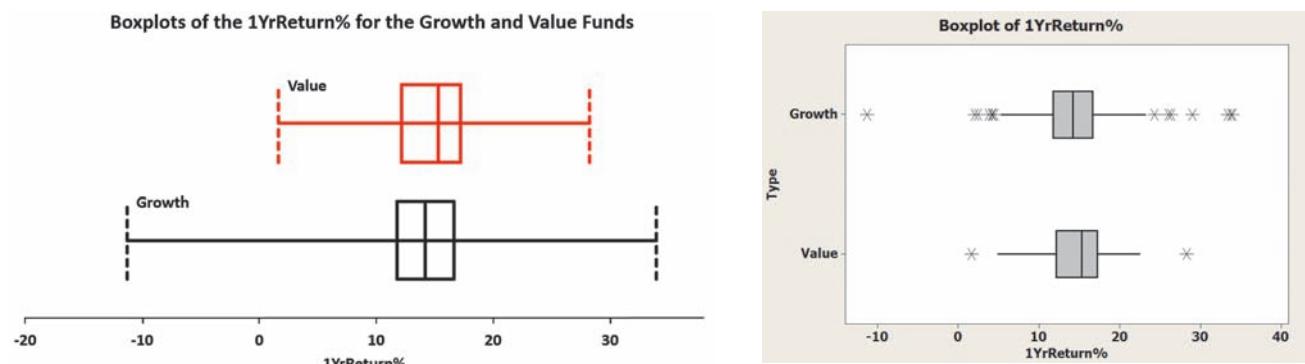
Boxplots of the One-Year Returns for the Growth and Value Funds

In the More Descriptive Choices scenario, you are interested in comparing the past performance of the growth and value funds from a sample of 316 funds. One measure of past performance is the one-year return percentage variable. Construct the boxplots for this variable for the growth and value funds.

SOLUTION Figure 3.4 contains the boxplots for the one-year return percentages for the growth and value funds. The five-number summary for the growth funds associated with these boxplots is $X_{\text{smallest}} = -11.28$, $Q_1 = 11.78$, median = 14.18, $Q_3 = 16.64$, and $X_{\text{largest}} = 33.98$. The five-number summary for the value funds associated with these boxplots is $X_{\text{smallest}} = 1.67$, $Q_1 = 12.17$, median = 15.3, $Q_3 = 17.23$, and $X_{\text{largest}} = 28.27$.

FIGURE 3.4

Excel and Minitab boxplots for the one-year return percentage variable



The lines, or whiskers, that extend from the Minitab central box each extend 1.5 times the interquartile range from the box. Beyond these ranges are the values considered to be outliers, plotted as asterisks.

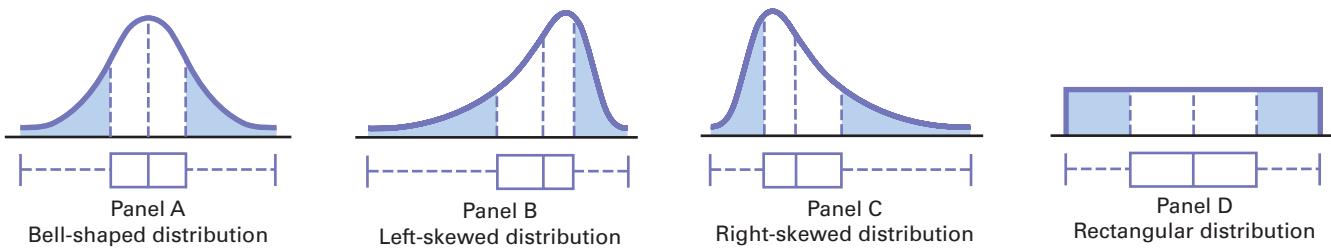
The median return, the quartiles, and the maximum returns are slightly higher for the value funds than for the growth funds. Both the growth and value funds are somewhat symmetrical, but the growth funds have a much larger range. These results are consistent with the statistics computed in Figure 3.2 on page 115.

Figure 3.5 demonstrates the relationship between the boxplot and the density curve for four different types of distributions. The area under each density curve is split into quartiles corresponding to the five-number summary for the boxplot.

The distributions in Panels A and D of Figure 3.5 are symmetrical. In these distributions, the mean and median are equal. In addition, the length of the left tail is equal to the length of the right tail, and the median line divides the box in half.

FIGURE 3.5

Boxplots and corresponding density curves for four distributions

**Student Tip**

A long tail on the left side of the boxplot indicates a left-skewed distribution. A long tail on the right side of the boxplot indicates a right-skewed distribution.

The distribution in Panel B of Figure 3.5 is left-skewed. The few small values distort the mean toward the left tail. For this left-skewed distribution, there is a heavy clustering of values at the high end of the scale (i.e., the right side); 75% of all values are found between the left edge of the box (Q_1) and the end of the right tail (X_{largest}). There is a long left tail that contains the smallest 25% of the values, demonstrating the lack of symmetry in this data set.

The distribution in Panel C of Figure 3.5 is right-skewed. The concentration of values is on the low end of the scale (i.e., the left side of the boxplot). Here, 75% of all values are found between the beginning of the left tail and the right edge of the box (Q_3). There is a long right tail that contains the largest 25% of the values, demonstrating the lack of symmetry in this data set.

Problems for Section 3.3

LEARNING THE BASICS

3.27 The following is a set of data from a sample of $n = 7$:

12 7 4 9 0 7 3

- Compute the first quartile (Q_1), the third quartile (Q_3), and the interquartile range.
- List the five-number summary.
- Construct a boxplot and describe its shape.
- Compare your answer in (c) with that from Problem 3.3 (d) on page 117. Discuss.

3.28 The following is a set of data from a sample of $n = 6$:

7 4 9 7 3 12

- Compute the first quartile (Q_1), the third quartile (Q_3), and the interquartile range.
- List the five-number summary.
- Construct a boxplot and describe its shape.
- Compare your answer in (c) with that from Problem 3.2 (d) on page 117. Discuss.

3.29 The following is a set of data from a sample of $n = 5$:

7 4 9 8 2

- Compute the first quartile (Q_1), the third quartile (Q_3), and the interquartile range.
- List the five-number summary.
- Construct a boxplot and describe its shape.
- Compare your answer in (c) with that from Problem 3.1 (d) on page 117. Discuss.

3.30 The following is a set of data from a sample of $n = 5$:

7 -5 -8 7 9

- Compute the first quartile (Q_1), the third quartile (Q_3), and the interquartile range.
- List the five-number summary.
- Construct a boxplot and describe its shape.
- Compare your answer in (c) with that from Problem 3.4 (d) on page 117. Discuss.

APPLYING THE CONCEPTS

3.31 The file **AccountingPartners** contains the number of partners in a cohort of rising accounting firms that have been tagged as “firms to watch.” The firms have the following numbers of partners:

25 18 23 16 16 14 22 17 9 32 22 12 18 9 25 9 28 14
22 18 22 16 12 31 17 12 14 16 30 12 12 10 24 12 11

Source: Data extracted from bit.ly/11asPHm.

- Compute the first quartile (Q_1), the third quartile (Q_3), and the interquartile range.
- List the five-number summary.
- Construct a boxplot and describe its shape.

3.32 The file **MarketPenetration** contains the market penetration value (i.e., the percentage of the country population that are users) for the 15 countries that lead the world in total number of Facebook users:

52.56 33.09 5.37 19.41 32.52 41.69 51.61 30.12
39.07 30.62 38.16 49.35 27.13 53.45 40.01

Source: Data extracted from www.socialbakers.com/facebook-statistics/.

- Compute the first quartile (Q_1), the third quartile (Q_3), and the interquartile range.
- List the five-number summary.
- Construct a boxplot and describe its shape.

3.33 The file **HotelAway** contains the average room price (in US\$) paid in 2012 by people of various nationalities while traveling away from their home country:

171 169 158 164 158 135 150 141

Source: Data extracted from bit.ly/XPtFOY.

- Compute the first quartile (Q_1), the third quartile (Q_3), and the interquartile range.
- List the five-number summary.
- Construct a boxplot and describe its shape.

3.34 The file **SUV** contains the overall MPG of 2013 small SUVs:

22 23 21 22 25 26 22 22 21
19 22 22 26 23 24 21 22

Source: Data extracted from “Ratings,” *Consumer Reports*, April 2013, pp. 34–36.

- Compute the first quartile (Q_1), the third quartile (Q_3), and the interquartile range.
- List the five-number summary.
- Construct a boxplot and describe its shape.

3.35 The file **CD Rate** contains the yields for one-year CDs and five-year CDs, for 23 banks in the United States, as of March 20, 2013.

Source: Data extracted from www.Bankrate.com, March 20, 2013.

For each type of account:

- Compute the first quartile (Q_1), the third quartile (Q_3), and the interquartile range.
- List the five-number summary.
- Construct a boxplot and describe its shape.

3.36 A bank branch located in a commercial district of a city has the business objective of developing an improved process for serving customers during the noon-to-1:00 P.M. lunch period. The waiting time, in minutes, is defined as the time the customer enters the line to when he or she reaches the teller window. Data are collected from a sample of 15 customers during this hour. The file **Bank1** contains the results, which are listed below:

4.21 5.55 3.02 5.13 4.77 2.34 3.54 3.20
4.50 6.10 0.38 5.12 6.46 6.19 3.79

Another bank branch, located in a residential area, is also concerned with the noon-to-1:00 P.M. lunch hour. The waiting times, in minutes, collected from a sample of 15 customers during this hour, are contained in the file **Bank2** and listed here:

9.66	5.90	8.02	5.79	8.73	3.82	8.01	8.35
10.49	6.68	5.64	4.08	6.17	9.91	5.47	

- List the five-number summaries of the waiting times at the two bank branches.
- Construct boxplots and describe the shapes of the distributions for the two bank branches.
- What similarities and differences are there in the distributions of the waiting times at the two bank branches?

3.4 Numerical Descriptive Measures for a Population

Sections 3.1 and 3.2 discuss the statistics that can be computed to describe the properties of central tendency and variation for a sample. When you collect data for an entire population (see Section 1.3), you compute and analyze population *parameters* for these properties, including the population mean, population variance, and population standard deviation.

To help illustrate these parameters, consider the population of stocks for the 10 companies in the Dow Jones Industrial Average (DJIA) that form the “Dogs of the Dow,” defined as the 10 stocks in the DJIA whose dividend is the highest fraction of their price in the previous year. (These stocks are used in an alternative investment scheme popularized by Michael O’Higgins.) Table 3.6 contains the 2012 one-year returns (excluding dividends) for the 10 “Dow Dog” stocks of 2011. These data, stored in **DowDogs**, will be used to illustrate the population parameters discussed in this section.

TABLE 3.6

One-Year Return
for the “Dogs of the
Dow”

Stock	One-Year Return	Stock	One-Year Return
AT&T	11.5	DuPont	-1.8
Verizon	7.9	Johnson & Johnson	6.9
Merck	8.6	Intel	-14.9
Pfizer	15.9	Procter & Gamble	1.8
General Electric	17.2	Kraft Foods Group	11.5

Source: Data extracted from 1Stock1.com.

The Population Mean

The **population mean** is the sum of the values in the population divided by the population size, N . This parameter, represented by the Greek lowercase letter mu, μ , serves as a measure of central tendency. Equation (3.13) defines the population mean.

POPULATION MEAN

The population mean is the sum of the values in the population divided by the population size, N .

$$\mu = \frac{\sum_{i=1}^N X_i}{N} \quad (3.13)$$

where

μ = population mean

X_i = i th value of the variable X

$\sum_{i=1}^N X_i$ = summation of all X_i values in the population

N = number of values in the population

To compute the mean one-year return for the population of “Dow Dog” stocks in Table 3.6, use Equation (3.13):

$$\begin{aligned}\mu &= \frac{\sum_{i=1}^N X_i}{N} \\ &= \frac{11.5 + 7.9 + 8.6 + 15.9 + 17.2 + (-1.8) + 6.9 + (-14.9) + 1.8 + 11.5}{10} \\ &= \frac{64.6}{10} = 6.46\end{aligned}$$

Thus, the mean one-year return for the “Dow Dog” stocks is 6.46.

The Population Variance and Standard Deviation

The population variance and the population standard deviation parameters measure variation in a population. The **population variance** is the sum of the squared differences around the population mean divided by the population size, N , and the **population standard deviation** is the square root of the population variance. In practice, you will most likely use the population standard deviation because, unlike the population variance, the standard deviation will always be a number expressed in the same units as the original population data.

The lowercase Greek letter sigma, σ , represents the population standard deviation, and sigma squared, σ^2 , represents the population variance. Equations (3.14) and (3.15) define these parameters. The denominators for the right-side terms in these equations use N and not the $(n - 1)$ term that is found in Equations (3.6) and (3.7) on page 109 that define the sample variance and standard deviation.

POPULATION VARIANCE

The population variance is the sum of the squared differences around the population mean divided by the population size, N :

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} \quad (3.14)$$

where

μ = population mean

X_i = i th value of the variable X

$\sum_{i=1}^N (X_i - \mu)^2$ = summation of all the squared differences between the X_i values and μ

POPULATION STANDARD DEVIATION

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}} \quad (3.15)$$

To compute the population variance for the data of Table 3.6, you use Equation (3.14):

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} \\ &= \frac{25.4016 + 2.0736 + 4.5796 + 89.1136 + 115.3476 +}{10} \\ &\quad = \frac{68.2276 + 0.1936 + 456.2496 + 21.7156 + 25.4016}{10} \\ &= \frac{808.8304}{10} = 80.8304\end{aligned}$$

From Equation (3.15), the population sample standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}} = \sqrt{\frac{808.3040}{10}} = 8.9906$$

Therefore, the typical percentage return differs from the mean of 6.46 by approximately 8.99. This large amount of variation suggests that the “Dow Dog” stocks produce results that differ greatly.

The Empirical Rule

In most data sets, a large portion of the values tend to cluster somewhere near the mean. In right-skewed data sets, this clustering occurs to the left of the mean—that is, at a value less than the mean. In left-skewed data sets, the values tend to cluster to the right of the mean—that is, greater than the mean. In symmetrical data sets, where the median and mean are the same, the values often tend to cluster around the median and mean, producing a normal distribution (discussed in Chapter 6).

The **empirical rule** states that for population data that form a normal distribution, the following are true:

- Approximately 68% of the values are within ± 1 standard deviation from the mean.
- Approximately 95% of the values are within ± 2 standard deviations from the mean.
- Approximately 99.7% of the values are within ± 3 standard deviations from the mean.

The empirical rule helps you examine variability in a population as well as identify outliers. The empirical rule implies that for normal distributions, only about 1 out of 20 values will be beyond 2 standard deviations from the mean in either direction. As a general rule, you can consider values not found in the interval $\mu \pm 2\sigma$ as potential outliers. The rule also implies that only about 3 in 1,000 will be beyond 3 standard deviations from the mean. Therefore, values not found in the interval $\mu \pm 3\sigma$ are almost always considered outliers.

EXAMPLE 3.15

Using the Empirical Rule

A population of 2-liter bottles of cola is known to have a mean fill-weight of 2.06 liters and a standard deviation of 0.02 liter. The population is known to be bell-shaped. Describe the distribution of fill-weights. Is it very likely that a bottle will contain less than 2 liters of cola?

SOLUTION

$$\begin{aligned}\mu \pm \sigma &= 2.06 \pm 0.02 = (2.04, 2.08) \\ \mu \pm 2\sigma &= 2.06 \pm 2(0.02) = (2.02, 2.10) \\ \mu \pm 3\sigma &= 2.06 \pm 3(0.02) = (2.00, 2.12)\end{aligned}$$

(continued)

Using the empirical rule, you can see that approximately 68% of the bottles will contain between 2.04 and 2.08 liters, approximately 95% will contain between 2.02 and 2.10 liters, and approximately 99.7% will contain between 2.00 and 2.12 liters. Therefore, it is highly unlikely that a bottle will contain less than 2 liters.

The Chebyshev Rule

For heavily skewed sets of data and data sets that do not appear to be normally distributed, you should use the Chebyshev rule instead of the empirical rule. The **Chebyshev rule** (see reference 2) states that for any data set, regardless of shape, the percentage of values that are found within distances of k standard deviations from the mean must be at least

$$\left(1 - \frac{1}{k^2}\right) \times 100\%$$

You can use this rule for any value of k greater than 1. For example, consider $k = 2$. The Chebyshev rule states that at least $[1 - (1/2)^2] \times 100\% = 75\%$ of the values must be found within ± 2 standard deviations of the mean.

The Chebyshev rule is very general and applies to any distribution. The rule indicates *at least* what percentage of the values fall within a given distance from the mean. However, if the data set is approximately bell-shaped, the empirical rule will more accurately reflect the greater concentration of data close to the mean. Table 3.7 compares the Chebyshev and empirical rules.

TABLE 3.7

How Data Vary Around the Mean

Section EG3.4 describes the **VE-Variability workbook** that allows you to use Excel to explore the empirical and Chebyshev rules.

Interval	% of Values Found in Intervals Around the Mean	
	Chebyshev (any distribution)	Empirical Rule (normal distribution)
$(\mu - \sigma, \mu + \sigma)$	At least 0%	Approximately 68%
$(\mu - 2\sigma, \mu + 2\sigma)$	At least 75%	Approximately 95%
$(\mu - 3\sigma, \mu + 3\sigma)$	At least 88.89%	Approximately 99.7%

EXAMPLE 3.16

Using the Chebyshev Rule

As in Example 3.15, a population of 2-liter bottles of cola is known to have a mean fill-weight of 2.06 liter and a standard deviation of 0.02 liter. However, the shape of the population is unknown, and you cannot assume that it is bell-shaped. Describe the distribution of fill-weights. Is it very likely that a bottle will contain less than 2 liters of cola?

SOLUTION

$$\begin{aligned}\mu \pm \sigma &= 2.06 \pm 0.02 = (2.04, 2.08) \\ \mu \pm 2\sigma &= 2.06 \pm 2(0.02) = (2.02, 2.10) \\ \mu \pm 3\sigma &= 2.06 \pm 3(0.02) = (2.00, 2.12)\end{aligned}$$

Because the distribution may be skewed, you cannot use the empirical rule. Using the Chebyshev rule, you cannot say anything about the percentage of bottles containing between 2.04 and 2.08 liters. You can state that at least 75% of the bottles will contain between 2.02 and 2.10 liters and at least 88.89% will contain between 2.00 and 2.12 liters. Therefore, between 0 and 11.11% of the bottles will contain less than 2 liters.

You can use these two rules to understand how data are distributed around the mean when you have sample data. With each rule, you use the value you computed for \bar{X} in place of μ and the value you computed for S in place of σ . The results you compute using the sample statistics are *approximations* because you used sample statistics (\bar{X}, S) and not population parameters (μ, σ).

Problems for Section 3.4

LEARNING THE BASICS

- 3.37** The following is a set of data for a population with $N = 10$:

7 5 11 8 3 6 2 1 9 8

- Compute the population mean.
- Compute the population standard deviation.

- 3.38** The following is a set of data for a population with $N = 10$:

7 5 6 6 4 8 6 9 3

- Compute the population mean.
- Compute the population standard deviation.

APPLYING THE CONCEPTS

- 3.39** The file **Tax** contains the quarterly sales tax receipts (in \$thousands) submitted to the comptroller of the Village of Fair Lake for the period ending March 2013 by all 50 business establishments in that locale:

10.3	11.1	9.6	9.0	14.5	13.0	6.7	11.0	8.4	10.3
8.0	11.2	7.3	5.3	12.5	8.0	11.8	8.7	10.6	9.5
11.1	10.2	11.1	9.9	9.8	11.6	15.1	12.5	6.5	7.5
10.0	12.9	9.2	10.0	12.8	12.5	9.3	10.4	12.7	10.5
9.3	11.5	10.7	11.6	7.8	10.5	7.6	10.1	8.9	8.6

- Compute the mean, variance, and standard deviation for this population.
- What percentage of the 50 businesses has quarterly sales tax receipts within $\pm 1, 2$, or ± 3 standard deviations of the mean?
- Compare your findings with what would be expected on the basis of the empirical rule. Are you surprised at the results in (b)?

- 3.40** Consider a population of 1,024 mutual funds that primarily invest in large companies. You have determined that μ , the mean one-year total percentage return achieved by all the funds, is 8.20 and that σ , the standard deviation, is 2.75.

- According to the empirical rule, what percentage of these funds is expected to be within ± 1 standard deviation of the mean?
- According to the empirical rule, what percentage of these funds is expected to be within ± 2 standard deviations of the mean?

- According to the Chebyshev rule, what percentage of these funds is expected to be within $\pm 1, \pm 2$, or ± 3 standard deviations of the mean?
- According to the Chebyshev rule, at least 93.75% of these funds are expected to have one-year total returns between what two amounts?

- 3.41** The file **CigaretteTax** contains the state cigarette tax (in \$) for each of the 50 states as of January 1, 2013.

- Compute the population mean and population standard deviation for the state cigarette tax.
- Interpret the parameters in (a).

-  **3.42** The file **Energy** contains the per capita energy consumption, in kilowatt-hours, for each of the 50 states and the District of Columbia during a recent year.

- Compute the mean, variance, and standard deviation for the population.
- What proportion of these states has per capita energy consumption within ± 1 standard deviation of the mean, within ± 2 standard deviations of the mean, and within ± 3 standard deviations of the mean?
- Compare your findings with what would be expected based on the empirical rule. Are you surprised at the results in (b)?
- Repeat (a) through (c) with the District of Columbia removed. How have the results changed?

- 3.43** Thirty companies comprise the DJIA. Just how big are these companies? One common method for measuring the size of a company is to use its market capitalization, which is computed by multiplying the number of stock shares by the price of a share of stock. On March 30, 2013, the market capitalization of these companies ranged from Alcoa's \$9.1 billion to ExxonMobil's \$403.7 billion. The entire population of market capitalization values is stored in **DowMarketCap**.

Source: Data extracted from money.cnn.com, March 30, 2013.

- Compute the mean and standard deviation of the market capitalization for this population of 30 companies.
- Interpret the parameters computed in (a).

3.5 The Covariance and the Coefficient of Correlation

In Section 2.5, you used scatter plots to visually examine the relationship between two numerical variables. This section presents two measures of the relationship between two numerical variables: the covariance and the coefficient of correlation.

The Covariance

The **covariance** measures the strength of the linear relationship between two numerical variables (X and Y). Equation (3.16) defines the **sample covariance**, and Example 3.17 illustrates its use.

SAMPLE COVARIANCE

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1} \quad (3.16)$$

EXAMPLE 3.17

Computing the Sample Covariance

In Figure 2.14 on page 66, you constructed a scatter plot that showed the relationship between the value and the annual revenue of the 30 NBA professional basketball teams (stored in **NBAValues**). Now, you want to measure the association between the annual revenue and value of a team by determining the sample covariance.

SOLUTION Table 3.8 provides the annual revenue and the value of the 30 teams.

TABLE 3.8

Revenues and Values for NBA Teams

Team Code	Revenue (\$millions)	Value (\$millions)	Team Code	Revenue (\$millions)	Value (\$millions)
ATL	99	316	MIA	150	625
BOS	143	730	MIL	87	312
BKN	84	530	MIN	96	364
CHA	93	315	NOH	100	340
CHI	162	800	NYK	243	1,100
CLE	128	434	OKC	127	475
DAL	137	685	ORL	126	470
DEN	110	427	PHI	107	418
DET	125	400	PHX	121	474
GSW	127	555	POR	117	457
HOU	135	568	SAC	96	525
IND	98	383	SAS	135	527
LAC	108	430	TOR	121	405
LAL	197	1,000	UTA	111	432
MEM	96	377	WAS	102	397

Figure 3.6 contains two worksheets that together compute the covariance for these data. From the result in cell B9 of the covariance worksheet, or by using Equation (3.16) directly (shown below), you determine that the covariance is 5,767.7:

$$\begin{aligned}\text{cov}(X, Y) &= \frac{167,263.3}{30 - 1} \\ &= 5,767.7\end{aligned}$$

FIGURE 3.6

Excel data and covariance worksheets for the revenue and value for the 30 NBA teams

A	B	C	D	
1	Revenue	Value	(X-XBar)	(Y-YBar)
2	99	316	-23.7000	-193.0333
3	143	730	20.3000	220.9667
4	84	530	-38.7000	20.9667
5	93	315	-29.7000	-194.0333
6	162	800	39.3000	290.9667
7	128	434	5.3000	-75.0333
8	137	685	14.3000	175.9667
9	110	427	-12.7000	-82.0333
10	125	400	2.3000	-109.0333
11	127	555	4.3000	45.9667
12	135	568	12.3000	58.9667
13	98	383	-24.7000	-126.0333
14	108	430	-14.7000	-79.0333
15	197	1000	74.3000	490.9667
16	96	377	-26.7000	-132.0333

A	B
1	Covariance Analysis of Revenue and Value
2	
3	Intermediate Calculations
4	XBar
5	YBar
6	$\Sigma(X-XBar)(Y-YBar)$
7	n-1
8	
9	Covariance
	5767.7000
	=COVARIANCE.S(DATA!A:A, DATA!B:B)

In Figure 3.6, the covariance worksheet illustration includes a list of formulas to the right of the cells in which they occur, a style used throughout the rest of this book.

The covariance has a major flaw as a measure of the linear relationship between two numerical variables. Because the covariance can have any value, you cannot use it to determine the relative strength of the relationship. In Example 3.17, you cannot tell whether the value 5,767.7 indicates a strong relationship or a weak relationship between revenue and value. To better determine the relative strength of the relationship, you need to compute the coefficient of correlation.

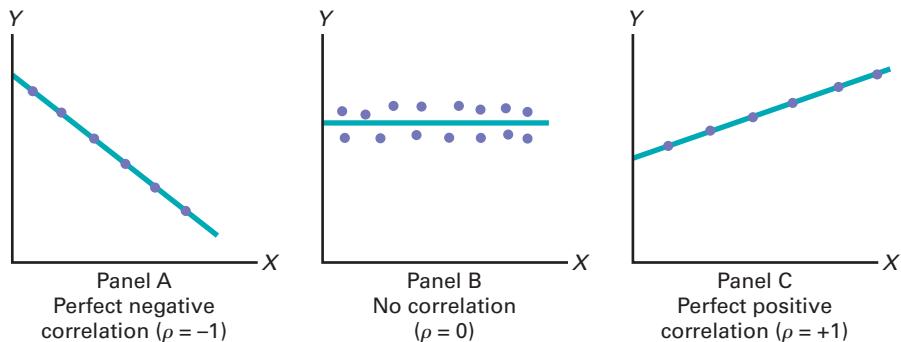
The Coefficient of Correlation

The **coefficient of correlation** measures the relative strength of a linear relationship between two numerical variables. The values of the coefficient of correlation range from -1 for a perfect negative correlation to $+1$ for a perfect positive correlation. *Perfect* in this case means that if the points were plotted on a scatter plot, all the points could be connected with a straight line.

When dealing with population data for two numerical variables, the Greek letter ρ (*rho*) is used as the symbol for the coefficient of correlation. Figure 3.7 illustrates three different types of association between two variables.

FIGURE 3.7

Types of association between variables



In Panel A of Figure 3.7, there is a perfect negative linear relationship between X and Y . Thus, the coefficient of correlation, ρ , equals -1 , and when X increases, Y decreases in a perfectly predictable manner. Panel B shows a situation in which there is no relationship between X and Y . In this case, the coefficient of correlation, ρ , equals 0 , and as X increases, there is no tendency for Y to increase or decrease. Panel C illustrates a perfect positive relationship where ρ equals $+1$. In this case, Y increases in a perfectly predictable manner when X increases.

Correlation alone cannot prove that there is a causation effect—that is, that the change in the value of one variable caused the change in the other variable. A strong correlation can be produced by chance; by the effect of a **lurking variable**—the third variable not considered in the calculation of the correlation; or by a cause-and-effect relationship. You would need to perform additional analysis to determine which of these three situations actually produced the correlation. Therefore, you can say that *causation implies correlation, but correlation alone does not imply causation.*

Equation (3.17) defines the **sample coefficient of correlation (r)**.

SAMPLE COEFFICIENT OF CORRELATION

$$r = \frac{\text{cov}(X, Y)}{S_X S_Y} \quad (3.17)$$

where

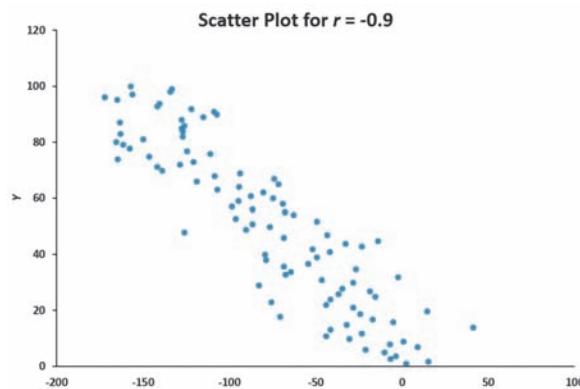
$$\begin{aligned} \text{cov}(X, Y) &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1} \\ S_X &= \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}} \\ S_Y &= \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}} \end{aligned}$$

When you have sample data, you can compute the sample coefficient of correlation, r . When using sample data, you are unlikely to have a sample coefficient of correlation of exactly $+1$, 0 , or -1 . Figure 3.8 presents scatter plots along with their respective sample coefficients of correlation, r , for six data sets, each of which contains $100 X$ and Y values.

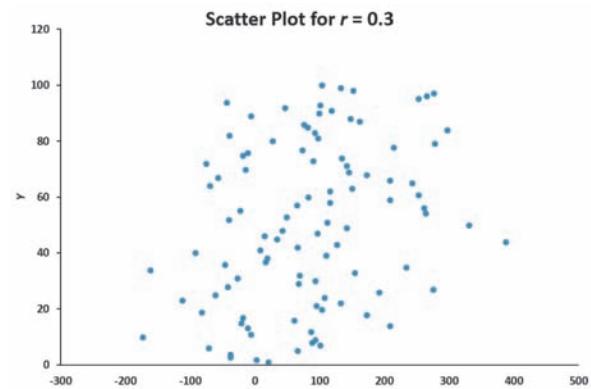
In Panel A, the coefficient of correlation, r , is -0.9 . You can see that for small values of X , there is a very strong tendency for Y to be large. Likewise, the large values of X tend to be paired with small values of Y . The data do not all fall on a straight line, so the association between X and Y cannot be described as perfect. The data in Panel B have a coefficient of correlation equal to -0.6 , and the small values of X tend to be paired with large values of Y . The linear relationship between X and Y in Panel B is not as strong as that in Panel A. Thus, the coefficient of correlation in Panel B is not as negative as that in Panel A. In Panel C, the linear relationship between X and Y is very weak, $r = -0.3$, and there is only a slight tendency for the small values of X to be paired with the large values of Y . Panels D through F depict data sets that have positive coefficients of correlation because small values of X tend to be paired with small values of Y , and large values of X tend to be associated with large values of Y . Panel D shows weak positive correlation, with $r = 0.3$. Panel E shows stronger positive correlation, with $r = 0.6$. Panel F shows very strong positive correlation, with $r = 0.9$.

FIGURE 3.8

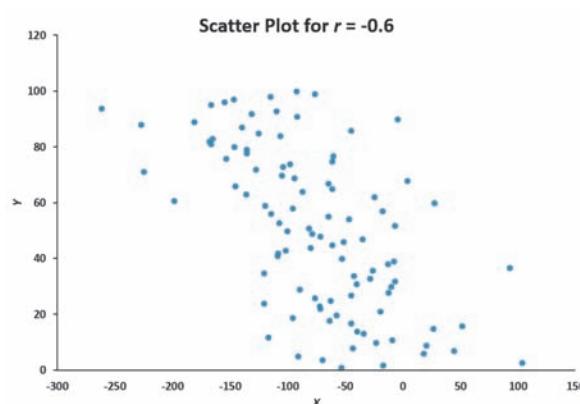
Six scatter plots and their sample coefficients of correlation, r



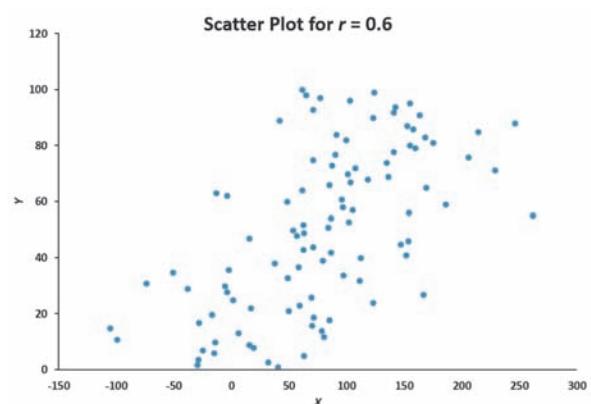
Panel A



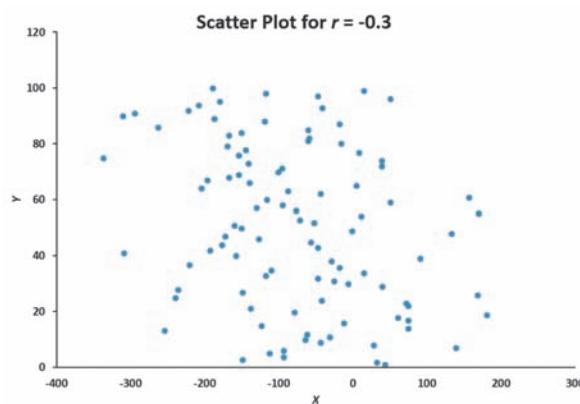
Panel D



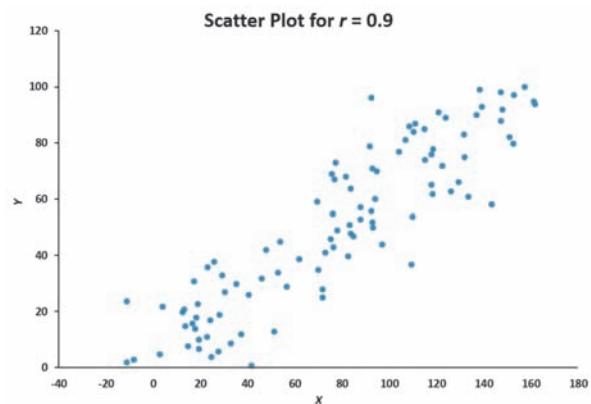
Panel B



Panel E



Panel C



Panel F

EXAMPLE 3.18**Computing the Sample Coefficient of Correlation**

In Example 3.17 on page 132, you computed the covariance of the revenue and value for the 30 NBA teams. Now, you want to measure the relative strength of a linear relationship between the revenue and value by determining the sample coefficient of correlation.

SOLUTION By using Equation (3.17) directly (shown below) or from cell B14 in the coefficient of correlation worksheet (shown in Figure 3.9), you determine that the sample coefficient of correlation is 0.9143:

$$\begin{aligned} r &= \frac{\text{cov}(X, Y)}{S_X S_Y} \\ &= \frac{5,767.70}{(33.2132)(189.9329)} \\ &= 0.9143 \end{aligned}$$

FIGURE 3.9

Excel worksheet to compute the sample coefficient of correlation between revenue and value

The Figure 3.9 worksheet uses the data worksheet shown in Figure 3.6 on page 133.

A	B
1	Coefficient of Correlation Analysis
2	
3	Intermediate Calculations
4	XBar 122.7000 =AVERAGE(DATA!A:A)
5	YBar 509.0333 =AVERAGE(DATA!B:B)
6	$\Sigma(X-X\bar{ })^2$ 31990.3000 =DEVSQ(DATA!A:A)
7	$\Sigma(Y-Y\bar{ })^2$ 1046160.9667 =DEVSQ(DATA!B:B)
8	$\Sigma(X-X\bar{ })(Y-Y\bar{ })$ 167263.3000 =SUMPRODUCT(DATA!C:C, DATA!D:D)
9	n-1 29 =COUNT(DATA!A:A) - 1
10	Covariance 5767.7000 =COVARIANCE.S(DATA!A:A, DATA!B:B)
11	S_x 33.2132 =SQRT(B6/B9)
12	S_y 189.9329 =SQRT(B7/B9)
13	
14	r 0.9143 =CORREL(DATA!A:A, DATA!B:B)

The value and revenue of the NBA teams are very highly correlated. The teams with the lowest revenues have the lowest values. The teams with the highest revenues have the highest values. This relationship is very strong, as indicated by the coefficient of correlation, $r = 0.9143$.

In general, you cannot assume that just because two variables are correlated, changes in one variable caused changes in the other variable. However, for this example, it makes sense to conclude that changes in revenue would tend to cause changes in the value of a team.

In summary, the coefficient of correlation indicates the linear relationship, or association, between two numerical variables. When the coefficient of correlation gets closer to +1 or -1, the linear relationship between the two variables is stronger. When the coefficient of correlation is near 0, little or no linear relationship exists. The sign of the coefficient of correlation indicates whether the data are positively correlated (i.e., the larger values of X are typically paired with the larger values of Y) or negatively correlated (i.e., the larger values of X are typically paired with the smaller values of Y). The existence of a strong correlation does not imply a causation effect. It only indicates the tendencies present in the data.

Problems for Section 3.5**LEARNING THE BASICS**

3.44 The following is a set of data from a sample of $n = 11$ items:

X	7	5	8	3	6	10	12	4	9	15	18
Y	21	15	24	9	18	30	36	12	27	45	54

- Compute the covariance.
- Compute the coefficient of correlation.
- How strong is the relationship between X and Y ? Explain.

APPLYING THE CONCEPTS

3.45 A study of 1,839 college students suggests a link between frequency of using Facebook and texting in class and grade point average. Students reporting a higher frequency of Facebook and texting use in class had lower grade point averages than students reporting a lower frequency of Facebook and texting use. (Source: Data extracted from “In-Class Multi-Tasking and Academic Programs.” *Computers in Human Computers*, 2012, dx.doi.org/10.1016/j.chb.2012.06.031).

- Does the study suggest that frequency of using Facebook and texting in class and grade point average are positively correlated or negatively correlated?
- Do you think that there might be a cause-and-effect relationship between frequency of using Facebook and texting in class and grade point average? Explain.

 **SELF Test** **3.46** The file **Cereals** lists the calories and sugar, in grams, in one serving of seven breakfast cereals:

Cereal	Calories	Sugar
Kellogg's All Bran	80	6
Kellogg's Corn Flakes	100	2
Wheaties	100	4
Nature's Path Organic Multigrain Flakes	110	4
Kellogg's Rice Krispies	130	4
Post Shredded Wheat Vanilla Almond	190	11
Kellogg's Mini Wheats	200	10

- Compute the covariance.
- Compute the coefficient of correlation.
- Which do you think is more valuable in expressing the relationship between calories and sugar—the covariance or the coefficient of correlation? Explain.
- Based on (a) and (b), what conclusions can you reach about the relationship between calories and sugar?

3.47 Movie companies need to predict the gross receipts of individual movies once a movie has debuted. The data, shown in the next column and stored in **PotterMovies**, are the first weekend gross, the U.S. gross, and the worldwide gross (in \$millions) of the eight Harry Potter movies:

- Compute the covariance between first weekend gross and U.S. gross, first weekend gross and worldwide gross, and U.S. gross and worldwide gross.
- Compute the coefficient of correlation between first weekend gross and U.S. gross, first weekend gross and worldwide gross, and U.S. gross and worldwide gross.

Title	First Weekend	U.S. Gross	Worldwide Gross
<i>Sorcerer's Stone</i>	90.295	317.558	976.458
<i>Chamber of Secrets</i>	88.357	261.988	878.988
<i>Prisoner of Azkaban</i>	93.687	249.539	795.539
<i>Goblet of Fire</i>	102.335	290.013	896.013
<i>Order of the Phoenix</i>	77.108	292.005	938.469
<i>Half-Blood Prince</i>	77.836	301.460	934.601
<i>Deathly Hallows – Part 1</i>	125.017	295.001	955.417
<i>Deathly Hallows – Part 2</i>	169.189	381.011	1,328.111

Source: Data extracted from www.the-numbers.com/interactive/comp-HarryPotter.php.

- Which do you think is more valuable in expressing the relationship between first weekend gross, U.S. gross, and worldwide gross—the covariance or the coefficient of correlation? Explain.
- Based on (a) and (b), what conclusions can you reach about the relationship between first weekend gross, U.S. gross, and worldwide gross?

3.48 College football is big business, with coaches' total pay and revenues, in millions of dollars. The file **College Football** contains the coaches' pay and revenues for college football at 105 of the 124 schools that are part of the Division I Football Bowl Subdivision.

Source: Data extracted from “College Football Coaches Continue to See Salary Explosion,” *USA Today*, November 20, 2012.

- Compute the covariance.
- Compute the coefficient of correlation.
- Based on (a) and (b), what conclusions can you reach about the relationship between coaches' total pay and revenues?

3.49 A Pew Research Center survey found that social networking is popular in many nations around the world. The file **GlobalSocialMedia** contains the level of social media networking (measured as the percentage of individuals polled who use social networking sites) and the GDP at purchasing power parity (PPP) per capita for each of 25 selected countries. (Data extracted from Pew Research Center, “Global Digital Communication: Texting, Social Networking Popular Worldwide,” updated February 29, 2012, via the link bit.ly/sNjsmq).

- Compute the covariance.
- Compute the coefficient of correlation.
- Based on (a) and (b), what conclusions can you reach about the relationship between the GDP and social media use?

3.6 Descriptive Statistics: Pitfalls and Ethical Issues

This chapter describes how a set of numerical data can be characterized by the statistics that measure the properties of central tendency, variation, and shape. In business, descriptive statistics such as the ones discussed in this chapter are frequently included in summary reports that are prepared periodically.

The volume of information available from online, broadcast, or print media has produced much skepticism in the minds of many about the objectivity of data. When you are reading information that contains descriptive statistics, you should keep in mind the quip often attributed to the famous nineteenth-century British statesman Benjamin Disraeli: “There are three kinds of lies: lies, damned lies, and statistics.”

For example, in examining statistics, you need to compare the mean and the median. Are they similar, or are they very different? Or is only the mean provided? The answers to these questions will help you determine whether the data are skewed or symmetrical and whether the median might be a better measure of central tendency than the mean. In addition, you should look to see whether the standard deviation or interquartile range for a very skewed set of data has been included in the statistics provided. Without this, it is impossible to determine the amount of variation that exists in the data.

Ethical considerations arise when you are deciding what results to include in a report. You should document both good and bad results. In addition, when making oral presentations and presenting written reports, you need to give results in a fair, objective, and neutral manner. Unethical behavior occurs when you selectively fail to report pertinent findings that are detrimental to the support of a particular position.

USING STATISTICS

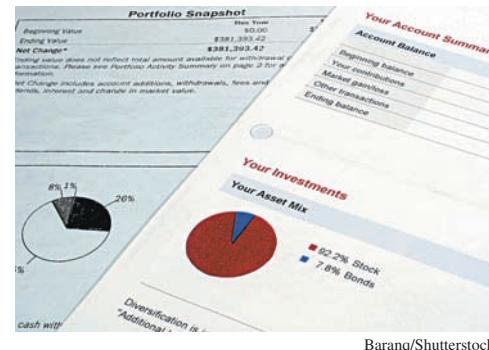
More Descriptive Choices, Revisited

In the More Descriptive Choices scenario, you were hired by the Choice Is Yours investment company to assist investors interested in stock mutual funds. A sample of 316 stock mutual funds included 227 growth funds and 89 value funds. By comparing these two categories, you were able to provide investors with valuable insights.

The one-year returns for both the growth funds and the value funds were symmetrical, as indicated by the boxplots (see Figures 3.4 and 3.5 on page 125). The descriptive statistics (see Figure 3.2 on page 115) allowed you to compare the central tendency, variability, and shape of the returns of the growth funds and the value funds. The mean indicated that the growth funds returned an average of 14.28, and the median indicated that half of the growth funds had returns of 14.18 or more. The value funds’ central tendencies were slightly higher than those of the growth funds—they had a mean of

14.70, and half the funds had one-year returns above 15.30. The growth funds showed

slightly more variability than the value funds, with a standard deviation of 5.0041 as compared to 4.4651. The kurtosis of growth funds was very positive, indicating a distribution that was much more peaked than a normal distribution. Although past performance is no assurance of future performance, the value funds slightly outperformed the growth funds in 2012. (You can examine other variables in [Retirement Funds](#) to see if the value funds outperformed the growth funds for the three-year period 2010–2012, for the 5-year period 2008–2012 and for the 10-year period 2003–2012.)



Barang/Shutterstock

SUMMARY

In this chapter and the previous chapter, you studied descriptive statistics—how you can organize data through tables, visualize data through charts, and how you can use various statistics to help analyze the data and reach conclusions. In Chapter 2, you organized data by constructing summary tables and visualized data by constructing bar and pie charts, histograms, and other charts. In this chapter, you learned how descriptive statistics such as

the mean, median, quartiles, range, and standard deviation describe the characteristics of central tendency, variability, and shape. In addition, you constructed boxplots to visualize the distribution of the data. You also learned how the coefficient of correlation describes the relationship between two numerical variables. All the methods of this chapter are summarized in Table 3.9.

TABLE 3.9

Chapter 3 Descriptive Statistics Methods

Type of Analysis	Methods
Central tendency	Mean, median, mode (Section 3.1)
Variation and shape	Quartiles, range, interquartile range, variance, standard deviation, coefficient of variation, Z scores, boxplot (Sections 3.2 through 3.4)
Describing the relationship between two numerical variables	Covariance, coefficient of correlation (Section 3.5)

You also learned several concepts about variation in data that will prove useful in later chapters. These concepts are:

- The greater the spread or dispersion of the data, the larger the range, variance, and standard deviation.
- The smaller the spread or dispersion of the data, the smaller the range, variance, and standard deviation.
- If the values are all the same (so that there is no variation in the data), the range, variance, and standard deviation will all equal zero.

- None of the measures of variation (the range, variance, and standard deviation) can ever be negative.

In the next chapter, the basic principles of probability are presented in order to bridge the gap between the subject of descriptive statistics and the subject of inferential statistics.

REFERENCES

1. Booker, J., and L. Ticknor. "A Brief Overview of Kurtosis." www.osti.gov/bridge/purl.cover.jsp?purl=/677174-zdulqk/webviewable/677174.pdf.
2. Kendall, M. G., A. Stuart, and J. K. Ord. *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*, 6th ed. New York: Oxford University Press, 1994.
3. Microsoft Excel 2013. Redmond, WA: Microsoft Corporation, 2012.
4. Minitab Release 16. State College, PA: Minitab, Inc., 2010.
5. Taleb, N. *The Black Swan*, 2nd ed. New York: Random House, 2010.

KEY EQUATIONS

Sample Mean

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (3.1)$$

Median

$$\text{Median} = \frac{n+1}{2} \text{ ranked value} \quad (3.2)$$

Geometric Mean

$$\bar{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n} \quad (3.3)$$

Geometric Mean Rate of Return

$$\bar{R}_G = [(1 + R_1) \times (1 + R_2) \times \cdots \times (1 + R_n)]^{1/n} - 1 \quad (3.4)$$

Range

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}} \quad (3.5)$$

Sample Variance

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \quad (3.6)$$

Sample Standard Deviation

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} \quad (3.7)$$

Coefficient of Variation

$$CV = \left(\frac{S}{\bar{X}} \right) 100\% \quad (3.8)$$

Z Score

$$Z = \frac{X - \bar{X}}{S} \quad (3.9)$$

First Quartile, Q_1

$$Q_1 = \frac{n+1}{4} \text{ ranked value} \quad (3.10)$$

Third Quartile, Q_3

$$Q_3 = \frac{3(n+1)}{4} \text{ ranked value} \quad (3.11)$$

Interquartile Range

$$\text{Interquartile range} = Q_3 - Q_1 \quad (3.12)$$

Population Mean

$$\mu = \frac{\sum_{i=1}^N X_i}{N} \quad (3.13)$$

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} \quad (3.14)$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}} \quad (3.15)$$

Sample Covariance

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \quad (3.16)$$

Sample Coefficient of Correlation

$$r = \frac{\text{cov}(X, Y)}{S_X S_Y} \quad (3.17)$$

KEY TERMS

arithmetic mean (mean) 102
 boxplot 124
 central tendency 102
 Chebyshev rule 130
 coefficient of correlation 133
 coefficient of variation (CV) 112
 covariance 132
 dispersion (spread) 107
 empirical rule 129
 five-number summary 123
 geometric mean 106
 interquartile range (midspread) 122
 kurtosis 115
 left-skewed 114
 lepokurtic 115
 lurking variable 134

mean (arithmetic mean) 102
 median 104
 midspread (interquartile range) 122
 mode 105
 outliers 113
 percentiles 121
 platykurtic 115
 population mean 127
 population standard deviation 128
 population variance 128
 Q_1 : first quartile 120
 Q_2 : second quartile 120
 Q_3 : third quartile 120
 quartiles 120
 range 107
 resistant measure 123

right-skewed 114
 sample coefficient of correlation (r) 134
 sample covariance 132
 sample mean 102
 sample standard deviation (S) 108
 sample variance (S^2) 108
 shape 114
 skewed 114
 skewness 114
 spread (dispersion) 107
 standard deviation 108
 sum of squares (SS) 108
 symmetrical 114
 variance 108
 variation 102
 Z score 113

CHECKING YOUR UNDERSTANDING

- 3.50** What are the properties of a set of numerical data?
- 3.51** What is meant by the property of central tendency?
- 3.52** What are the differences among the mean, median, and mode, and what are the advantages and disadvantages of each?
- 3.53** How do you interpret the first quartile, median, and third quartile?
- 3.54** What is meant by the property of variation?
- 3.55** What does the Z score measure?
- 3.56** What are the differences among the various measures of variation, such as the range, interquartile range, variance, standard

deviation, and coefficient of variation, and what are the advantages and disadvantages of each?

- 3.57** How does the empirical rule help explain the ways in which the values in a set of numerical data cluster and distribute?
- 3.58** How do the empirical rule and the Chebyshev rule differ?
- 3.59** What is meant by the property of shape?
- 3.60** What is the difference between the arithmetic mean and the geometric mean?
- 3.61** What is the difference between skewness and kurtosis?
- 3.62** How do the covariance and the coefficient of correlation differ?

CHAPTER REVIEW PROBLEMS

3.63 The American Society for Quality (ASQ) conducted a salary survey of all its members. ASQ members work in all areas of manufacturing and service-related institutions, with a common theme of an interest in quality. For the survey, emails were sent to 54,337 members, and 6,093 valid responses were received. Manager and quality engineer were the most frequently reported job titles among the valid responses. Master Black Belt, a person who takes a leadership role as the keeper of the Six Sigma process (see Section 19.6) and Green Belt, someone who works on Six Sigma projects part time, were among the other job titles cited. Descriptive statistics concerning salaries for these four titles are given in the following table:

Job Title	Sample Size	Standard		Mean	Median
		Minimum	Maximum		
Green Belt	26	36,000	135,000	23,399	74,173
Manager	1,387	38,000	732,000	27,906	91,878
Quality Engineer	766	25,000	185,000	21,933	79,575
Master Black Belt	79	28,000	185,000	27,474	112,946
					110,000

Source: Data extracted from M. Hansen "Facing Tight Times," *Quality Progress*, December 2012, p. 29.

Compare the salaries of Green Belts, managers, quality engineers, and Master Black Belts.

3.64 In certain states, savings banks are permitted to sell life insurance. The approval process consists of underwriting, which includes a review of the application, a medical information bureau check, possible requests for additional medical information and medical exams, and a policy compilation stage, in which the policy pages are generated and sent to the bank for delivery. The ability to deliver approved policies to customers in a timely manner is critical to the profitability of this service to the bank. Using the Define, Collect, Organize, Visualize, and Analyze steps first discussed in Chapter 2, you define the variable of interest as the total processing time in days. You collect the data by selecting a random sample of 27 approved policies during a period of one month. You organize the data collected in a worksheet and store them in **Insurance**:

73 19 16 64 28 28 31 90 60 56 31 56 22 18
45 48 17 17 17 91 92 63 50 51 69 16 17

- a. Compute the mean, median, first quartile, and third quartile.
- b. Compute the range, interquartile range, variance, standard deviation, and coefficient of variation.
- c. Construct a boxplot. Are the data skewed? If so, how?
- d. What would you tell a customer who enters the bank to purchase this type of insurance policy and asks how long the approval process takes?

3.65 One of the major measures of the quality of service provided by an organization is the speed with which it responds to customer complaints. A large family-held department store selling furniture and flooring, including carpet, had undergone a major expansion in the past several years. In particular, the flooring department had expanded from 2 installation crews to an installation supervisor, a measurer, and 15 installation crews. The business objective of the company was to reduce the time between when a complaint is received and when it is resolved. During a recent year, the company received 50 complaints concerning carpet installation. The data from the 50 complaints, organized in **Furniture**, represent the number of days between the receipt of a complaint and the resolution of the complaint:

54	5	35	137	31	27	152	2	123	81	74	27	11
19	126	110	110	29	61	35	94	31	26	5	12	4
165	32	29	28	29	26	25	1	14	13	13	10	5
27	4	52	30	22	36	26	20	23	33	68		

- a. Compute the mean, median, first quartile, and third quartile.
- b. Compute the range, interquartile range, variance, standard deviation, and coefficient of variation.
- c. Construct a boxplot. Are the data skewed? If so, how?
- d. On the basis of the results of (a) through (c), if you had to tell the president of the company how long a customer should expect to wait to have a complaint resolved, what would you say? Explain.

3.66 A manufacturing company produces steel housings for electrical equipment. The main component part of the housing is a steel trough that is made of a 14-gauge steel coil. It is produced using a 250-ton progressive punch press with a wipe-down operation and two 90-degree forms placed in the flat steel to make the trough. The distance from one side of the form to the other is critical because of weatherproofing in outdoor applications. The company requires that the width of the trough be between 8.31 inches and 8.61 inches. Data are collected from a sample of 49 troughs and stored in **Trough**, which contains these widths of the troughs, in inches:

8.312 8.343 8.317 8.383 8.348 8.410 8.351 8.373 8.481 8.422
8.476 8.382 8.484 8.403 8.414 8.419 8.385 8.465 8.498 8.447
8.436 8.413 8.489 8.414 8.481 8.415 8.479 8.429 8.458 8.462
8.460 8.444 8.429 8.460 8.412 8.420 8.410 8.405 8.323 8.420
8.396 8.447 8.405 8.439 8.411 8.427 8.420 8.498 8.409

- a. Compute the mean, median, range, and standard deviation for the width. Interpret these measures of central tendency and variability.
- b. List the five-number summary.
- c. Construct a boxplot and describe its shape.
- d. What can you conclude about the number of troughs that will meet the company's requirement of troughs being between 8.31 and 8.61 inches wide?

3.67 The manufacturing company in Problem 3.66 also produces electric insulators. If the insulators break when in use, a short circuit is likely to occur. To test the strength of the insulators, destructive testing is carried out to determine how much force is required to break the insulators. Force is measured by observing how many pounds must be applied to an insulator before it breaks. Data are collected from a sample of 30 insulators. The file **Force** contains the strengths, as follows:

1,870 1,728 1,656 1,610 1,634 1,784 1,522 1,696 1,592 1,662
 1,866 1,764 1,734 1,662 1,734 1,774 1,550 1,756 1,762 1,866
 1,820 1,744 1,788 1,688 1,810 1,752 1,680 1,810 1,652 1,736

- Compute the mean, median, range, and standard deviation for the force needed to break the insulators.
- Interpret the measures of central tendency and variability in (a).
- Construct a boxplot and describe its shape.
- What can you conclude about the strength of the insulators if the company requires a force of at least 1,500 pounds before breakage?

3.68 Data were collected on the typical cost of dining at American-cuisine restaurants within a 1-mile walking distance of a hotel located in a large city. The file **Bundle** contains the typical cost (a per transaction cost in \$) as well as a Bundle score, a measure of overall popularity and customer loyalty, for each of 40 selected restaurants. (Data extracted from www.bundle.com via the link [on-msn.com/MnlBxo](#).)

- For each variable, compute the mean, median, first quartile, and third quartile.
- For each variable, compute the range, interquartile range, variance, standard deviation, and coefficient of variation.
- For each variable, construct a boxplot. Are the data skewed? If so, how?
- Compute the coefficient of correlation between Bundle score and typical cost.
- What conclusions can you reach concerning Bundle score and typical cost?

3.69 A quality characteristic of interest for a tea-bag-filling process is the weight of the tea in the individual bags. If the bags are underfilled, two problems arise. First, customers may not be able to brew the tea to be as strong as they wish. Second, the company may be in violation of the truth-in-labeling laws. For this product, the label weight on the package indicates that, on average, there are 5.5 grams of tea in a bag. If the mean amount of tea in a bag exceeds the label weight, the company is giving away product. Getting an exact amount of tea in a bag is problematic because of variation in the temperature and humidity inside the factory, differences in the density of the tea, and the extremely fast filling operation of the machine (approximately 170 bags per minute). The file **Teabags** contains these weights, in grams, of a sample of 50 tea bags produced in one hour by a single machine:

5.65 5.44 5.42 5.40 5.53 5.34 5.54 5.45 5.52 5.41
 5.57 5.40 5.53 5.54 5.55 5.62 5.56 5.46 5.44 5.51
 5.47 5.40 5.47 5.61 5.53 5.32 5.67 5.29 5.49 5.55
 5.77 5.57 5.42 5.58 5.58 5.50 5.32 5.50 5.53 5.58
 5.61 5.45 5.44 5.25 5.56 5.63 5.50 5.57 5.67 5.36

- Compute the mean, median, first quartile, and third quartile.
- Compute the range, interquartile range, variance, standard deviation, and coefficient of variation.
- Interpret the measures of central tendency and variation within the context of this problem. Why should the company producing the tea bags be concerned about the central tendency and variation?
- Construct a boxplot. Are the data skewed? If so, how?
- Is the company meeting the requirement set forth on the label that, on average, there are 5.5 grams of tea in a bag? If you were in charge of this process, what changes, if any, would you try to make concerning the distribution of weights in the individual bags?

3.70 The manufacturer of Boston and Vermont asphalt shingles provides its customers with a 20-year warranty on most of its products. To determine whether a shingle will last as long as the warranty period, accelerated-life testing is conducted at the manufacturing plant. Accelerated-life testing exposes a shingle to the stresses it would be subject to in a lifetime of normal use via an experiment in a laboratory setting that takes only a few minutes to conduct. In this test, a shingle is repeatedly scraped with a brush for a short period of time, and the shingle granules removed by the brushing are weighed (in grams). Shingles that experience low amounts of granule loss are expected to last longer in normal use than shingles that experience high amounts of granule loss. In this situation, a shingle should experience no more than 0.8 gram of granule loss if it is expected to last the length of the warranty period. The file **Granule** contains a sample of 170 measurements made on the company's Boston shingles and 140 measurements made on Vermont shingles.

- List the five-number summaries for the Boston shingles and for the Vermont shingles.
- Construct side-by-side boxplots for the two brands of shingles and describe the shapes of the distributions.
- Comment on the ability of each type of shingle to achieve a granule loss of 0.8 gram or less.
- What conclusions can you reach about the cost of a meal at city and suburban restaurants?

3.71 The file **Restaurants** contains the cost per meal and the ratings of 50 city and 50 suburban restaurants on their food, décor, and service (and their summated ratings). Complete the following for the urban and suburban restaurants:

Source: Data extracted from *Zagat Survey 2013 New York City Restaurants* and *Zagat Survey 2012–2013 Long Island Restaurants*.

- Construct the five-number summary of the cost of a meal.
- Construct a boxplot of the cost of a meal. What is the shape of the distribution?
- Compute and interpret the correlation coefficient of the summated rating and the cost of a meal.
- What conclusions can you reach about the cost of a meal at city and suburban restaurants?

3.72 The file **Protein** contains calories, protein, and cholesterol of popular protein foods (fresh red meats, poultry, and fish).

Source: U.S. Department of Agriculture.

- Compute the correlation coefficient between calories and protein.
- Compute the correlation coefficient between calories and cholesterol.
- Compute the correlation coefficient between protein and cholesterol.
- Based on the results of (a) through (c), what conclusions can you reach concerning calories, protein, and cholesterol?

3.73 The file **HotelPrices** contains the prices in British pounds of a room at two-star, three-star, and four-star hotels in cities around the world in 2012. (Data extracted from bit.ly/Q0qxe4.) Complete the following for two-star, three-star, and four-star hotels:

- Compute the mean, median, first quartile, and third quartile.
- Compute the range, interquartile range, variance, standard deviation, and coefficient of variation.
- Interpret the measures of central tendency and variation within the context of this problem.
- Construct a boxplot. Are the data skewed? If so, how?
- Compute the covariance between the average price at two-star and three-star hotels, between two-star and four-star hotels, and between three-star and four-star hotels.
- Compute the coefficient of correlation between the average price at two-star and three-star hotels, between two-star and four-star hotels, and between three-star and four-star hotels.
- Which do you think is more valuable in expressing the relationship between the average price of a room at two-star, three-star, and four-star hotels—the covariance or the coefficient of correlation? Explain.
- Based on (f), what conclusions can you reach about the relationship between the average price of a room at two-star, three-star, and four-star hotels?

3.74 The file **PropertyTaxes** contains the property taxes per capita for the 50 states and the District of Columbia.

- Compute the mean, median, first quartile, and third quartile.
- Compute the range, interquartile range, variance, standard deviation, and coefficient of variation.
- Construct a boxplot. Are the data skewed? If so, how?
- Based on the results of (a) through (c), what conclusions can you reach concerning property taxes per capita for each state and the District of Columbia?

3.75 The file **CEO-Compensation** includes the total compensation (in \$millions) of CEOs of 170 large public companies and the investment return in 2012.

Source: Data extracted from “CEO Pay Rockets as Economy, Stocks Recover,” *USA Today*, March 27, 2013, p. 1B.

- Compute the mean, median, first quartile, and third quartile.
- Compute the range, interquartile range, variance, standard deviation, and coefficient of variation.
- Construct a boxplot. Are the data skewed? If so, how?
- Based on the results of (a) through (c), what conclusions can you reach concerning the total compensation (in \$millions) of CEOs?
- Compute the correlation coefficient between compensation and the investment return in 2012.
- What conclusions can you reach from the results of (e)?

3.76 311 is Chicago’s web and phone portal for government information and nonemergency services. 311 serves as a comprehensive one-stop shop for residents, visitors, and business owners; therefore, it is critical that 311 representatives answer calls and respond to requests in a timely and accurate fashion. The target response time for answering 311 calls is 45 seconds. Agent abandonment rate is one of several call center metrics tracked by 311 officials. This metric tracks the percentage of callers who hang up after the target response time of 45 seconds has elapsed. The file **311CallCenter** contains the agent abandonment rate for 22 weeks of call center operation during the 7:00 A.M.–3:00 P.M. shift.

- Compute the mean, median, first quartile, and third quartile.
- Compute the range, interquartile range, variance, standard deviation, and coefficient of variation.

- Construct a boxplot. Are the data skewed? If so, how?
- Compute the correlation coefficient between day and agent abandonment rate.
- Based on the results of (a) through (c), what conclusions might you reach concerning 311 call center performance operation?

3.77 How much time do Americans living in or near cities spend waiting in traffic, and how much does waiting in traffic cost them per year? The file **Congestion** includes this cost for 31 cities. (Source: Data extracted from “The High Cost of Congestion,” *Time*, October 17, 2011, p. 18.) For the time Americans living in or near cities spend waiting in traffic and the cost of waiting in traffic per year:

- Compute the mean, median, first quartile, and third quartile.
- Compute the range, interquartile range, variance, standard deviation, and coefficient of variation.
- Construct a boxplot. Are the data skewed? If so, how?
- Compute the correlation coefficient between the time spent sitting in traffic and the cost of sitting in traffic.
- Based on the results of (a) through (c), what conclusions might you reach concerning the time spent waiting in traffic and the cost of waiting in traffic.

3.78 How do the average credit scores of people living in various American cities differ? The file **Credit Scores** is an ordered array of the average credit scores of people living in 143 American cities. (Data extracted from usat.ly/17a1fA6)

- Compute the mean, median, first quartile, and third quartile.
- Compute the range, interquartile range, variance, standard deviation, and coefficient of variation.
- Construct a boxplot. Are the data skewed? If so, how?
- Based on the results of (a) through (c), what conclusions might you reach concerning the average credit scores of people living in various American cities?

3.79 You are planning to study for your statistics examination with a group of classmates, one of whom you particularly want to impress. This individual has volunteered to use Microsoft Excel to generate the needed summary information, tables, and charts for a data set that contains several numerical and categorical variables assigned by the instructor for study purposes. This person comes over to you with the printout and exclaims, “I’ve got it all—the means, the medians, the standard deviations, the boxplots, the pie charts—for all our variables. The problem is, some of the output looks weird—like the boxplots for gender and for major and the pie charts for grade point average and for height. Also, I can’t understand why Professor Szabat said we can’t get the descriptive stats for some of the variables; I got them for everything! See, the mean for height is 68.23, the mean for grade point average is 2.76, the mean for gender is 1.50, the mean for major is 4.33.” What is your reply?

REPORT WRITING EXERCISES

3.80 The file **DomesticBeer** contains the percentage alcohol, number of calories per 12 ounces, and number of carbohydrates (in grams) per 12 ounces for 152 of the best-selling domestic beers in the United States. (Data extracted from bit.ly/17H3Ct, March 20, 2013.) Write a report that includes a complete descriptive evaluation of each of the numerical variables—percentage of alcohol, number of calories per 12 ounces, and number of carbohydrates (in grams) per 12 ounces. Append to your report all appropriate tables, charts, and numerical descriptive measures.

CASES FOR CHAPTER 3

Managing Ashland MultiComm Services

For what variable in the Chapter 2 “Managing Ashland MultiComm Services” case (see page 82) are numerical descriptive measures needed?

1. For the variable you identify, compute the appropriate numerical descriptive measures and construct a boxplot.

2. For the variable you identify, construct a graphical display. What conclusions can you reach from this other plot that cannot be made from the boxplot?
3. Summarize your findings in a report that can be included with the task force’s study.

Digital Case

Apply your knowledge about the proper use of numerical descriptive measures in this continuing Digital Case from Chapter 2.

Open **EndRunGuide.pdf**, the EndRun Financial Services “Guide to Investing.” Reexamine EndRun’s supporting data for the “More Winners Than Losers” and “The Big Eight Difference” and then answer the following:

1. Can descriptive measures be computed for any variables? How would such summary statistics support EndRun’s

claims? How would those summary statistics affect your perception of EndRun’s record?

2. Evaluate the methods EndRun used to summarize the results presented on the “Customer Survey Results” page. Is there anything you would do differently to summarize these results?
3. Note that the last question of the survey has fewer responses than the other questions. What factors may have limited the number of responses to that question?

CardioGood Fitness

Return to the CardioGood Fitness case first presented on page 83. Using the data stored in **CardioGoodFitness**:

1. Compute descriptive statistics to create a customer profile for each CardioGood Fitness treadmill product line.

2. Write a report to be presented to the management of CardioGood Fitness, detailing your findings.

More Descriptive Choices Follow-up

Follow up the Using Statistics Revisited section on page 138 by computing descriptive statistics to analyze the differences in 3-year return percentages, 5-year return percentages, and 10-year return percentages for the sample of 316 retirement

funds stored in **Retirement Funds**. In your analysis, examine differences between the growth and value funds as well as the differences among the small, mid-cap, and large market cap funds.

Clear Mountain State Student Surveys

1. The student news service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students who attend CMSU. They create and distribute a survey of 14 questions and receive responses from 62 undergraduates (stored in **UndergradSurvey**). For each numerical variable included in the survey, compute all the appropriate descriptive statistics and write a report summarizing your conclusions.

2. The dean of students at CMSU has learned about the undergraduate survey and has decided to undertake a similar survey for graduate students at CMSU. She creates and distributes a survey of 14 questions and receives responses from 44 graduate students (stored in **GradSurvey**). For each numerical variable included in the survey, compute all the appropriate descriptive statistic and write a report summarizing your conclusions.

CHAPTER 3 EXCEL GUIDE

EG3.1 CENTRAL TENDENCY

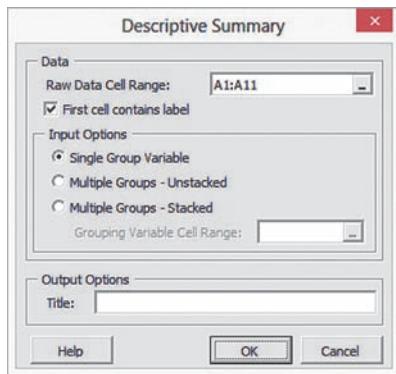
The Mean, Median, and Mode

Key Technique Use the **AVERAGE(variable cell range)**, **MEDIAN(variable cell range)**, and **MODE(variable cell range)** functions to compute these measures.

Example Compute the mean, median, and mode for the sample of getting-ready times introduced in Section 3.1.

PHStat Use Descriptive Summary.

For the example, open to the **DATA worksheet** of the **Times workbook**. Select **PHStat → Descriptive Statistics → Descriptive Summary**. In the procedure's dialog box (shown below):



1. Enter A1:A11 as the **Raw Data Cell Range** and check **First cell contains label**.
2. Click **Single Group Variable**.
3. Enter a **Title** and click **OK**.

PHStat inserts a new worksheet that contains various measures of central tendency, variation, and shape discussed in Sections 3.1 and 3.2. This worksheet is similar to the CompleteStatistics worksheet of the Descriptive workbook.

In-Depth Excel Use the CentralTendency worksheet of the Descriptive workbook as a model.

For the example, open the **Times workbook** and insert a new worksheet (see Section EG.4 on page 10) and:

1. Enter a title in cell **A1**.
2. Enter **Get-Ready Times** in cell **B3**, **Mean** in cell **A4**, **Median** in cell **A5**, and **Mode** in cell **A6**.
3. Enter the formula **=AVERAGE(DATA!A:A)** in cell **B4**, the formula **=MEDIAN(DATA!A:A)** in cell **B5**, and the formula **=MODE(DATA!A:A)** in cell **B6**.

For these functions, the *variable cell range* includes the name of the **DATA worksheet** because the data being summarized appears on the separate **DATA worksheet**.

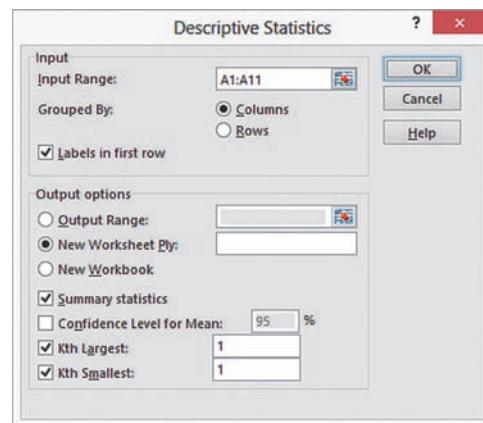
Analysis ToolPak Use Descriptive Statistics.

For the example, open to the **DATA worksheet** of the **Times workbook** and:

1. Select **Data → Data Analysis**.
2. In the Data Analysis dialog box, select **Descriptive Statistics** from the **Analysis Tools** list and then click **OK**.

In the Descriptive Statistics dialog box (shown below):

3. Enter A1:A11 as the **Input Range**. Click **Columns** and check **Labels in first row**.
4. Click **New Worksheet Ply** and check **Summary statistics**, **Kth Largest**, and **Kth Smallest**.
5. Click **OK**.



The ToolPak inserts a new worksheet that contains various measures of central tendency, variation, and shape discussed in Sections 3.1 and 3.2.

The Geometric Mean

Key Technique Use the **GEOMEAN((1 + (R1)),(1 + (R2)),...,(1 + (Rn))) – 1** function to compute the geometric mean rate of return.

Example Compute the geometric mean rate of return in the Russell 2000 Index for the two years as shown in Example 3.4 on page 107.

In-Depth Excel Enter the formula **=GEOMEAN((1 + (-0.055)),(1 + (0.146))) – 1** in any cell.

EG3.2 VARIATION and SHAPE

The Range

Key Technique Use the **MIN(variable cell range)** and **MAX(variable cell range)** functions to help compute the range.

Example Compute the range for the sample of getting-ready times first introduced in Section 3.1.

PHStat Use **Descriptive Summary** (see Section EG3.1).

In-Depth Excel Use the **Range worksheet** of the **Descriptive workbook** as a model.

For the example, open the worksheet implemented for the example in the *In-Depth Excel* “The Mean, Median, and Mode” instructions.

Enter **Minimum** in cell **A7**, **Maximum** in cell **A8**, and **Range** in cell **A9**. Enter the formula $=\text{MIN}(\text{DATA!A:A})$ in cell **B7**, the formula $=\text{MAX}(\text{DATA!A:A})$ in cell **B8**, and the formula $=\text{B8} - \text{B7}$ in cell **B9**.

The Variance, Standard Deviation, Coefficient of Variation, and Z Scores

Key Technique Use the **VAR.S(*variable cell range*)** and **STDEV.S(*variable cell range*)** functions to compute the sample variation and the sample standard deviation, respectively. Use the **AVERAGE** and **STDEV.S** functions for the coefficient of variation. Use the **STANDARDIZE(*value, mean, standard deviation*)** function to compute Z scores.

Example Compute the variance, standard deviation, coefficient of variation, and Z scores for the sample of getting-ready times first introduced in Section 3.1.

PHStat Use **Descriptive Summary** (see Section EG3.1).

In-Depth Excel Use the **Variation** and **ZScores worksheets** of the **Descriptive workbook** as models.

For the example, open to the worksheet implemented for the earlier examples. Enter **Variance** in cell **A10**, **Standard Deviation** in cell **A11**, and **Coeff. of Variation** in cell **A12**. Enter the formula $=\text{VAR.S}(\text{DATA!A:A})$ in cell **B10**, the formula $=\text{STDEV.S}(\text{DATA!A:A})$ in cell **B11**, and the formula $=\text{B11}/\text{AVERAGE}(\text{DATA!A:A})$ in cell **B12**. If you previously entered the formula for the mean in cell A4 using the Section EG3.1 *In-Depth Excel* instructions, enter the simpler formula $=\text{B11}/\text{B4}$ in cell **B12**. Right-click cell **B12** and click **Format Cells** in the shortcut menu. In the **Number** tab of the Format Cells dialog box, click **Percentage** in the **Category** list, enter **2** as the **Decimal places**, and click **OK**.

To compute the Z scores, copy the DATA worksheet. In the new, copied worksheet, enter **Z Score** in cell **B1**. Enter the formula $=\text{STANDARDIZE}(\text{A2}, \text{Variation!B$4}, \text{Variation!B$11})$ in cell **B2** and copy the formula down through row 11. If you use an Excel version older than Excel 2010, enter **Variation_Older!B\$4** and **Variation_Older!B\$11** as the cell references in the formula.

Analysis ToolPak Use **Descriptive Statistics** (see Section EG3.1). This procedure does not compute Z scores.

Shape: Skewness and Kurtosis

Key Technique Use the **SKEW(*variable cell range*)** and the **KURT(*variable cell range*)** functions to compute these measures.

Example Compute the skewness and kurtosis for the sample of getting-ready times first introduced in Section 3.1.

PHStat Use **Descriptive Summary** (see Section EG3.1).

In-Depth Excel Use the **Shape worksheet** of the **Descriptive workbook** as a model.

For the example, open to the worksheet implemented for the earlier examples. Enter **Skewness** in cell **A13** and **Kurtosis** in cell **A14**. Enter the formula $=\text{SKEW}(\text{DATA!A:A})$ in cell **B13** and the formula $=\text{KURT}(\text{DATA!A:A})$ in cell **B14**. Then format cells B13 and B14 for four decimal places.

Analysis ToolPak Use **Descriptive Statistics** (see Section EG3.1).

EG3.3 EXPLORING NUMERICAL DATA

Quartiles

Key Technique Use the **MEDIAN**, **COUNT**, **SMALL**, **INT**, **FLOOR**, and **CEILING** functions in combination with the **IF** decision-making function to compute the quartiles. To apply the rules of Section 3.3, avoid using any of the Excel quartile functions to compute the first and third quartiles.

Example Compute the quartiles for the sample of getting-ready times first introduced in Section 3.1.

PHStat Use **Boxplot** (discussed later on page 147).

In-Depth Excel Use the **COMPUTE worksheet** of the **Quartiles workbook** as a model.

For the example, the COMPUTE worksheet already computes the quartiles for the getting-ready times. To compute the quartiles for another set of data, paste the data into **column A** of the **DATA worksheet**, overwriting the existing getting-ready times.

Open to the **COMPUTE_FORMULAS worksheet** to examine the formulas and read the **SHORT TAKES** for Chapter 3 for an extended discussion of the formulas in the worksheet.

The workbook uses the older **QUARTILE(*variable cell range, quartile number*)** function and not the newer **QUARTILE.EXC** function for reasons explained in Appendix Section F.3. Both the older and newer functions use rules that differ from the Section 3.3 rules to compute quartiles. To compare the results using these newer functions, open to the **COMPARE worksheet**.

The Interquartile Range

Key Technique Use a formula to subtract the first quartile from the third quartile.

Example Compute the interquartile range for the sample of getting-ready times first introduced in Section 3.1.

In-Depth Excel Use the **COMPUTE worksheet** of the **Quartiles workbook** (introduced in the previous section) as a model.

For the example, the interquartile range is already computed in cell B19 using the formula $=\text{B18} - \text{B16}$.

The Five-Number Summary and the Boxplot

Key Technique Plot a series of line segments on the same chart to construct a boxplot. (Excel chart types do not include boxplots.)

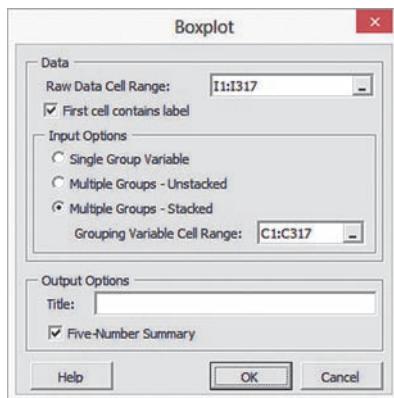
Example Compute the five-number summary and construct the boxplots of the one-year return percentage variable for the growth and value funds used in Example 3.14 on page 125.

PHStat Use Boxplot.

For the example, open to the **DATA worksheet** of the **Retirement Funds workbook**. Select **PHStat → Descriptive Statistics → Boxplot**. In the procedure's dialog box (shown below):

1. Enter **I1:I317** as the **Raw Data Cell Range** and check **First cell contains label**.
2. Click **Multiple Groups - Stacked** and enter **C1:C317** as the **Grouping Variable Cell Range**.
3. Enter a **Title**, check **Five-Number Summary**, and click **OK**.

The boxplot appears on its own chart sheet, separate from the worksheet that contains the five-number summary.



In-Depth Excel Use the worksheets of the **Boxplot workbook** as templates.

For the example, use the **PLOT_DATA worksheet** which already shows the five-number summary and boxplot for the value funds. To compute the five-number summary and construct a boxplot for the growth funds, copy the growth funds from **column A** of the **UNSTACKED worksheet** of the **Retirement Funds workbook** and paste into **column A** of the **DATA worksheet** of the **Boxplot workbook**.

For other problems, use the **PLOT_SUMMARY worksheet** as the template if the five-number summary has already been determined; otherwise, paste your unsummarized data into column A of the **DATA worksheet** and use the **PLOT_DATA worksheet** as was done for the example.

The worksheets creatively misuse Excel line charting features to construct a boxplot. Read the **SHORT TAKES** for Chapter 3 for an explanation of this “misuse.”

EG3.4 NUMERICAL DESCRIPTIVE MEASURES for a POPULATION

The Population Mean, Population Variance, and Population Standard Deviation

Key Technique Use **AVERAGE(variable cell range)**, **VAR.P(variable cell range)**, and **STDEV.P(variable cell range)** to compute these measures.

Example Compute the population mean, population variance, and population standard deviation for the “Dow Dogs” population data of Table 3.6 on page 127.

In-Depth Excel Use the **Parameters workbook** as a model. For the example, the **COMPUTE worksheet** of the **Parameters workbook** already computes the three population parameters for the “Dow Dogs.” If you use an Excel version older than Excel 2010, use the **COMPUTE_OLDER worksheet**.

The Empirical Rule and the Chebyshev Rule

Use the **COMPUTE worksheet** of the **VE-Variability workbook** to explore the effects of changing the mean and standard deviation on the ranges associated with ± 1 standard deviation, ± 2 standard deviations, and ± 3 standard deviations from the mean. Change the mean in cell **B4** and the standard deviation in cell **B5** and then note the updated results in rows 9 through 11.

EG3.5 THE COVARIANCE and the COEFFICIENT of CORRELATION

The Covariance

Key Technique Use the **COVARIANCE.S(variable 1 cell range, variable 2 cell range)** function to compute this measure.

Example Compute the sample covariance for the NBA team revenue and value shown in Figure 3.6 on page 133.

In-Depth Excel Use the **Covariance workbook** as a model.

For the example, the revenue and value have already been placed in columns A and B of the **DATA worksheet** and the **COMPUTE worksheet** displays the computed covariance in cell **B9**. For other problems, paste the data for two variables into columns A and B of the **DATA worksheet**, overwriting the revenue and value data.

Read the **SHORT TAKES** for Chapter 3 for an explanation of the formulas found in the **DATA** and **COMPUTE** worksheets. If you use an Excel version older than Excel 2010, use the **COMPUTE_OLDER** worksheet that computes the covariance without using the **COVARIANCE.S** function that was introduced in Excel 2010.

The Coefficient of Correlation

Key Technique Use the **CORREL(variable 1 cell range, variable 2 cell range)** function to compute this measure.

Example Compute the coefficient of correlation for the NBA team revenue and value data of Example 3.18 on page 136.

In-Depth Excel Use the **Correlation workbook** as a model.

For the example, the revenue and value have already been placed in columns A and B of the **DATA worksheet** and the **COMPUTE worksheet** displays the coefficient of correlation in cell **B14**. For other problems, paste the data for two variables into columns A and B of the **DATA worksheet**, overwriting the revenue and value data.

The **COMPUTE worksheet** that uses the **COVARIANCE.S** function to compute the covariance (see the previous section) and also uses the **DEVSQ**, **COUNT**, and **SUMPRODUCT** functions discussed in Appendix F. Open to the **COMPUTE_FORMULAS worksheet** to examine the use of all these functions.

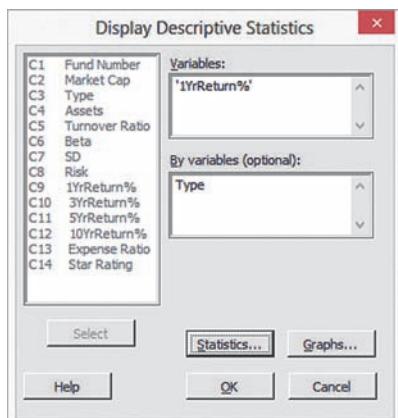
CHAPTER 3 MINITAB GUIDE

MG3.1 CENTRAL TENDENCY

The Mean, Median, and Mode

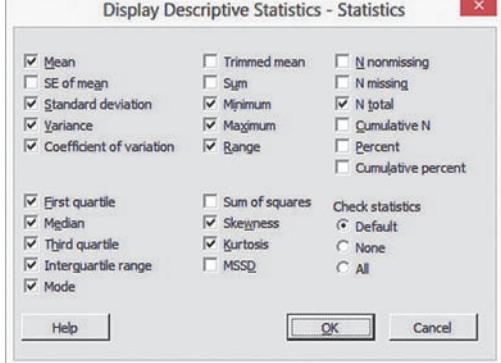
Use **Descriptive Statistics** to compute the mean, the median, the mode, and selected measures of variation and shape. For example, to create results similar to Figure 3.2 on page 115 that presents descriptive statistics of the one-year return percentage variable for the growth and value funds, open to the **Retirement Funds worksheet**. Select **Stat → Basic Statistics → Display Descriptive Statistics**. In the Display Descriptive Statistics dialog box (shown below):

1. Double-click **C9 1YrReturn%** in the variables list to add '**1YrReturn%**' to the **Variables** box and then press **Tab**.
2. Double-click **C3 Type** in the variables list to add **Type** to the **By variables (optional)** box.
3. Click **Statistics**.



In the Display Descriptive Statistics - Statistics dialog box (shown below):

4. Check **Mean, Standard deviation, Variance, Coefficient of variation, First quartile, Median, Third quartile, Interquartile range, Mode, Minimum, Maximum, Range, Skewness, Kurtosis, and N total**.
5. Click **OK**.



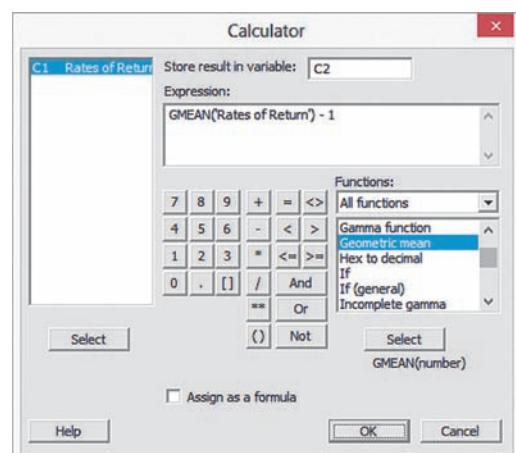
6. Back in the Display Descriptive Statistics dialog box, click **OK**.

The Geometric Mean

Use **Calculator** to compute the geometric mean or the geometric mean rate of return. For example, to compute the geometric mean

rate of return for Example 3.4 on page 107, open to the **Investments worksheet**. Select **Calc → Calculator**. In the Calculator dialog box (shown below):

1. Enter **C2** in the **Store result in variable** box and then press **Tab**. (C2 is the first empty column on the worksheet and the result will be placed in row 1 of column C2.)
2. Double-click **Geometric mean** in the **Functions** scrollable list to add **GMEAN(number)** to the **Expression** box.
3. Double-click **C1 Rates of Return** in the variables list to alter the expression to **GMEAN('Rates of Return')**. (If you prefer, you can directly edit the expression as part of the next step.)
4. Edit the expression so that it reads **GMEAN('Rates of Return') - 1**.
5. Click **OK**.



MG3.2 VARIATION and SHAPE

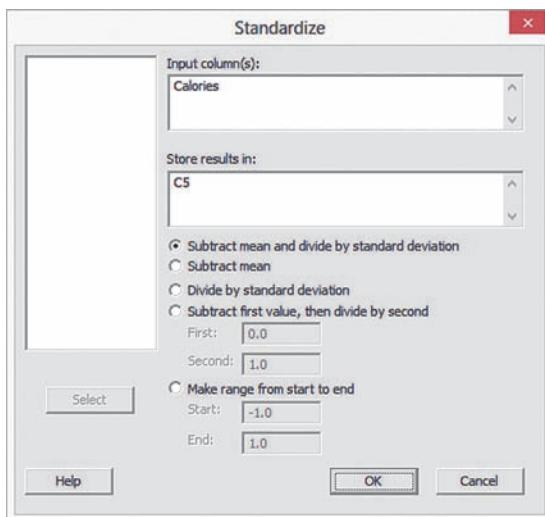
The Range, Variance, Standard Deviation, and Coefficient of Variation

Use **Descriptive Statistics** to compute these measures of variation and shape. The instructions in Section MG3.1 for computing the mean, median, and mode also compute these measures.

Z Scores

Use **Standardize** to compute Z scores. For example, to compute the Table 3.4 Z scores shown on page 114, open to the **CEREALS worksheet**. Select **Calc → Standardize**. In the Standardize dialog box (shown on page 149):

1. Double-click **C2 Calories** in the variables list to add **Calories** to the **Input column(s)** box and press **Tab**.
2. Enter **C5** in the **Store results in** box. (C5 is the first empty column on the worksheet and the Z scores will be placed in column C5.)
3. Click **Subtract mean and divide by standard deviation**.
4. Click **OK**.
5. In the new column C5, enter **Z Scores** as the name of the column.



Shape

Use **Descriptive Statistics** to compute skewness and kurtosis. The instructions in Section MG3.1 for computing the mean, median, and mode also compute these measures.

MG3.3 EXPLORING NUMERICAL DATA

Quartiles, the Interquartile Range, and the Five-Number Summary

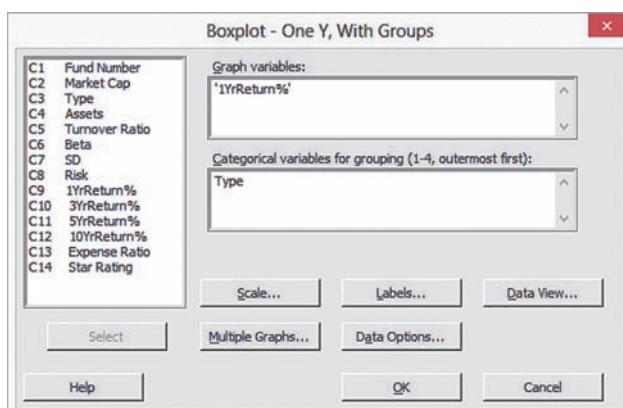
Use **Descriptive Statistics** to compute these measures. The instructions in Section MG3.1 for computing the mean, median, and mode also compute these measures.

The Boxplot

Use Boxplot.

For example, to create the Figure 3.4 boxplots on page 125, open to the **Retirement Funds worksheet**. Select **Graph → Boxplot**. In the Boxplots dialog box:

1. Click **With Groups** in the **One Y** gallery and then click **OK**.
2. Double-click **C9 1YrReturn%** in the variables list to add '**1YrReturn%**' to the **Graph variables** box and then press **Tab**.
3. Double-click **C3 Type** in the variables list to add **Type** in the **Categorical variables** box.
4. Click **OK**.



In the boxplot created, pausing the mouse pointer over the boxplot reveals a number of measures, including the quartiles. For problems that involve single-group data, click **Simple** in the **One Y** gallery in step 1.

To rotate the boxplots 90 degrees (as was done in Figure 3.5), replace step 4 with these steps 4 through 6:

4. Click **Scale**.
5. In the **Axes and Ticks** tab of the **Boxplot-Scale** dialog box, check **Transpose value and category scales** and click **OK**.
6. Back in the **Boxplot-One Y, With Groups** dialog box, click **OK**.

MG3.4 NUMERICAL DESCRIPTIVE MEASURES for a POPULATION

The Population Mean, Population Variance, and Population Standard Deviation

Minitab does not contain commands that compute these population parameters directly.

The Empirical Rule and the Chebyshev Rule

Manually compute the values needed to apply these rules using the statistics computed in the Section MG3.1 instructions.

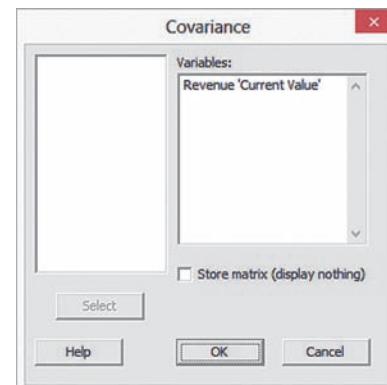
MG3.5 THE COVARIANCE and the COEFFICIENT of CORRELATION

The Covariance

Use Covariance.

For example, to compute the covariance for Example 3.17 on page 132, open to the **NBAValues worksheet**. Select **Stat → Basic Statistics → Covariance**. In the Covariance dialog box (shown below):

1. Double-click **C3 Revenue** in the variables list to add **Revenue** to the **Variables** box.
2. Double-click **C4 Current Value** in the variables list to add '**Current Value**' to the **Variables** box.



3. Click **OK**.

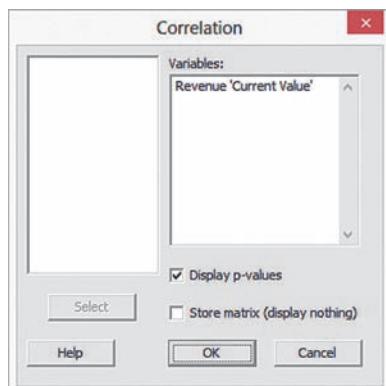
In the table of numbers produced, the covariance is the number that appears in the cell position that is the intersection of the two variables (the lower-left cell).

The Coefficient of Correlation

Use **Correlation**.

For example, to compute the coefficient of correlation for Example 3.18 on page 136, open to the **NBAValues worksheet**. Select **Stat → Basic Statistics → Correlation**. In the Correlation dialog box (shown in the right column):

1. Double-click **C3 Revenue** in the variables list to add **Revenue** to the **Variables** box.
2. Double-click **C4 Current Value** in the variables list to add '**Current Value**' to the **Variables** box.
3. Click **OK**.



CHAPTER

4

Basic Probability

CONTENTS

- 4.1 Basic Probability Concepts
- 4.2 Conditional Probability
- 4.3 Bayes' Theorem

THINK ABOUT THIS: Divine Providence and Spam

- 4.4 Counting Rules
- 4.5 Ethical Issues and Probability

USING STATISTICS: Possibilities at M&R Electronics World, Revised

CHAPTER 4 EXCEL GUIDE

CHAPTER 4 MINITAB GUIDE

OBJECTIVES

- To understand basic probability concepts
- To learn about conditional probability
- To use Bayes' theorem to revise probabilities
- To learn various counting rules

USING STATISTICS

Possibilities at M&R Electronics World

As the marketing manager for M&R Electronics World, you are analyzing the results of an intent-to-purchase study. The heads of 1,000 households were asked about their intentions to purchase a large-screen HDTV (one that has a screen size of at least 50 inches) sometime during the next 12 months. As a follow-up, you plan to survey the same people 12 months later to see whether they purchased a television. For households that did purchase a large-screen HDTV, you would like to know whether the television they purchased had a faster refresh rate (240 Hz or higher) or a standard refresh rate (60 or 120 Hz), whether they also purchased a streaming media box in the past 12 months, and whether they were satisfied with their purchase of the large-screen HDTV.

You plan to use the results of this survey to form a new marketing strategy that will enhance sales and better target those households likely to purchase multiple or more expensive products. What questions can you ask in this survey? How can you express the relationships among the various intent-to-purchase responses of individual households?



Shock/Fotolia

The principles of probability help bridge the worlds of descriptive statistics and inferential statistics. Probability principles are the foundation for the probability distribution, the concept of mathematical expectation, and the binomial, Poisson, and hypergeometric distributions, topics that are discussed in Chapter 5. In this chapter, you will learn about probability to answer questions such as the following:

- What is the probability that a household is planning to purchase a large-screen HDTV in the next year?
- What is the probability that a household will actually purchase a large-screen HDTV?
- What is the probability that a household is planning to purchase a large-screen HDTV and actually purchases the television?
- Given that the household is planning to purchase a large-screen HDTV, what is the probability that the purchase is made?
- Does knowledge of whether a household *plans* to purchase the television change the likelihood of predicting whether the household *will* purchase the television?
- What is the probability that a household that purchases a large-screen HDTV will purchase a television with a faster refresh rate?
- What is the probability that a household that purchases a large-screen HDTV with a faster refresh rate will also purchase a streaming media box?
- What is the probability that a household that purchases a large-screen HDTV will be satisfied with the purchase?

With answers to questions such as these, you can begin to form a marketing strategy. You can consider whether to target households that have indicated an intent to purchase or to focus on selling televisions that have faster refresh rates or both. You can also explore whether households that purchase large-screen HDTVs with faster refresh rates can be easily persuaded to also purchase streaming media boxes.

4.1 Basic Probability Concepts

Student Tip

Remember, a probability cannot be negative or greater than 1.

What is meant by the word *probability*? A **probability** is the numerical value representing the chance, likelihood, or possibility that a particular event will occur, such as the price of a stock increasing, a rainy day, a defective product, or the outcome five dots in a single toss of a die. In all these instances, the probability involved is a proportion or fraction whose value ranges between 0 and 1, inclusive. An event that has no chance of occurring (the **impossible event**) has a probability of 0. An event that is sure to occur (the **certain event**) has a probability of 1.

There are three types of probability:

- *A priori*
- Empirical
- Subjective

In the simplest case, where each outcome is equally likely, the chance of occurrence of the event is defined in Equation (4.1).

PROBABILITY OF OCCURRENCE

$$\text{Probability of occurrence} = \frac{X}{T} \quad (4.1)$$

where

X = number of ways in which the event occurs
 T = total number of possible outcomes

In *a priori* probability, the probability of an occurrence is based on prior knowledge of the process involved. Consider a standard deck of cards that has 26 red cards and 26 black cards. The probability of selecting a black card is $26/52 = 0.50$ because there are $X = 26$ black cards and $T = 52$ total cards. What does this probability mean? If each card is replaced after it is selected, does it mean that 1 out of the next 2 cards selected will be black? No, because you cannot say for certain what will happen on the next several selections. However, you can say that in the long run, if this selection process is continually repeated, the proportion of black cards selected will approach 0.50. Example 4.1 shows another example of computing an *a priori* probability.

EXAMPLE 4.1

Finding A Priori Probabilities

A standard six-sided die has six faces. Each face of the die contains either one, two, three, four, five, or six dots. If you roll a die, what is the probability that you will get a face with five dots?

SOLUTION Each face is equally likely to occur. Because there are six faces, the probability of getting a face with five dots is $1/6$.

The preceding examples use the *a priori* probability approach because the number of ways the event occurs and the total number of possible outcomes are known from the composition of the deck of cards or the faces of the die.

In the **empirical probability** approach, the probabilities are based on observed data, not on prior knowledge of a process. Surveys are often used to generate empirical probabilities. Examples of this type of probability are the proportion of individuals in the Using Statistics scenario who actually purchase large-screen HDTVs, the proportion of registered voters who prefer a certain political candidate, and the proportion of students who have part-time jobs. For example, if you take a survey of students, and 60% state that they have part-time jobs, then there is a 0.60 probability that an individual student has a part-time job.

The third approach to probability, **subjective probability**, differs from the other two approaches because subjective probability differs from person to person. For example, the development team for a new product may assign a probability of 0.60 to the chance of success for the product, while the president of the company may be less optimistic and assign a probability of 0.30. The assignment of subjective probabilities to various outcomes is usually based on a combination of an individual's past experience, personal opinion, and analysis of a particular situation. Subjective probability is especially useful in making decisions in situations in which you cannot use *a priori* probability or empirical probability.

Events and Sample Spaces

The basic elements of probability theory are the individual outcomes of a variable under study. You need the following definitions to understand probabilities.

Student Tip

Events are represented by letters of the alphabet.

EVENT

Each possible outcome of a variable is referred to as an **event**. A **simple event** is described by a single characteristic.

For example, when you toss a coin, the two possible outcomes are heads and tails. Each of these represents a simple event. When you roll a standard six-sided die in which the six faces of the die contain either one, two, three, four, five, or six dots, there are six possible simple events. An event can be any one of these simple events, a set of them, or a subset of all of them. For example, the event of an *even number of dots* consists of three simple events (i.e., two, four, or six dots).

Student Tip

The key word when describing a joint event is *and*.

JOINT EVENT

A **joint event** is an event that has two or more characteristics.

Getting two heads when you toss a coin twice is an example of a joint event because it consists of heads on the first toss and heads on the second toss.

COMPLEMENT

The **complement** of event A (represented by the symbol A') includes all events that are not part of A .

The complement of a head is a tail because that is the only event that is not a head. The complement of five dots on a die is not getting five dots. Not getting five dots consists of getting one, two, three, four, or six dots.

SAMPLE SPACE

The collection of all the possible events is called the **sample space**.

The sample space for tossing a coin consists of heads and tails. The sample space when rolling a die consists of one, two, three, four, five, and six dots. Example 4.2 demonstrates events and sample spaces.

EXAMPLE 4.2**Events and Sample Spaces****TABLE 4.1**

Purchase Behavior for Large-screen HDTVs

The Using Statistics scenario on page 151 concerns M&R Electronics World. Table 4.1 presents the results of the sample of 1,000 households in terms of purchase behavior for large-screen HDTVs.

PLANNED TO PURCHASE	ACTUALLY PURCHASED		
	Yes	No	Total
Yes	200	50	250
No	100	650	750
Total	300	700	1,000

What is the sample space? Give examples of simple events and joint events.

SOLUTION The sample space consists of the 1,000 respondents. Simple events are “planned to purchase,” “did not plan to purchase,” “purchased,” and “did not purchase.” The complement of the event “planned to purchase” is “did not plan to purchase.” The event “planned to purchase and actually purchased” is a joint event because in this joint event, the respondent must plan to purchase the television *and* actually purchase it.

Contingency Tables and Venn Diagrams

There are several ways in which you can view a particular sample space. One way involves using a **contingency table** (see Section 2.1) such as the one displayed in Table 4.1. You get the values in the cells of the table by subdividing the sample space of 1,000 households according to whether someone planned to purchase and actually purchased a large-screen HDTV. For example, 200 of the respondents planned to purchase a large-screen HDTV and subsequently did purchase the large-screen HDTV.

A second way to present the sample space is by using a **Venn diagram**. This diagram graphically represents the various events as “unions” and “intersections” of circles. Figure 4.1 presents a typical Venn diagram for a two-variable situation, with each variable having only two events (A and A' , B and B'). The circle on the left (the red one) represents all events that are part of A .

FIGURE 4.1

Venn diagram for events A and B

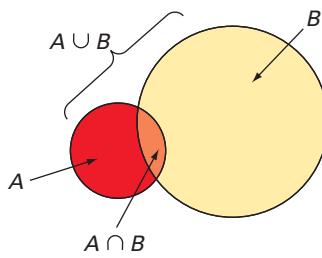
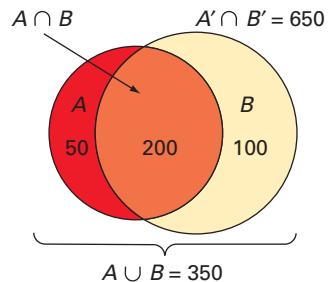


FIGURE 4.2

Venn diagram for the M&R Electronics World example



The circle on the right (the yellow one) represents all events that are part of B . The area contained within circle A and circle B (center area) is the intersection of A and B (written as $A \cap B$), since it is part of A and also part of B . The total area of the two circles is the union of A and B (written as $A \cup B$) and contains all outcomes that are just part of event A , just part of event B , or part of both A and B . The area in the diagram outside of $A \cup B$ contains outcomes that are neither part of A nor part of B .

You must define A and B in order to develop a Venn diagram. You can define either event as A or B , as long as you are consistent in evaluating the various events. For the large-screen HDTV example, you can define the events as follows:

A = planned to purchase B = actually purchased

A' = did not plan to purchase B' = did not actually purchase

In drawing the Venn diagram (see Figure 4.2), you must determine the value of the intersection of A and B so that the sample space can be divided into its parts. $A \cap B$ consists of all 200 households who planned to purchase and actually purchased a large-screen HDTV. The remainder of event A (planned to purchase) consists of the 50 households who planned to purchase a large-screen HDTV but did not actually purchase one. The remainder of event B (actually purchased) consists of the 100 households who did not plan to purchase a large-screen HDTV but actually purchased one. The remaining 650 households represent those who neither planned to purchase nor actually purchased a large-screen HDTV.

Simple Probability

Now you can answer some of the questions posed in the Using Statistics scenario. Because the results are based on data collected in a survey (refer to Table 4.1), you can use the empirical probability approach.

As stated previously, the most fundamental rule for probabilities is that they range in value from 0 to 1. An impossible event has a probability of 0, and an event that is certain to occur has a probability of 1.

Simple probability refers to the probability of occurrence of a simple event, $P(A)$. A simple probability in the Using Statistics scenario is the probability of planning to purchase

a large-screen HDTV. How can you determine the probability of selecting a household that planned to purchase a large-screen HDTV? Using Equation (4.1) on page 152:

$$\text{Probability of occurrence} = \frac{X}{T}$$

$$\begin{aligned} P(\text{Planned to purchase}) &= \frac{\text{Number who planned to purchase}}{\text{Total number of households}} \\ &= \frac{250}{1,000} = 0.25 \end{aligned}$$

Thus, there is a 0.25 (or 25%) chance that a household planned to purchase a large-screen HDTV.

Example 4.3 illustrates another application of simple probability.

EXAMPLE 4.3

Computing the Probability That the Large-Screen HDTV Purchased Had a Faster Refresh Rate

TABLE 4.2

Purchase Behavior Regarding Purchasing a Faster Refresh Rate Television and a Streaming Media Box

In the Using Statistics follow-up survey, additional questions were asked of the 300 households that actually purchased large-screen HDTVs. Table 4.2 indicates the consumers' responses to whether the television purchased had a faster refresh rate and whether they also purchased a streaming media box in the past 12 months.

Find the probability that if a household that purchased a large-screen HDTV is randomly selected, the television purchased had a faster refresh rate.

REFRESH RATE OF TELEVISION PURCHASED	STREAMING MEDIA BOX		
	Yes	No	Total
Faster	38	42	80
Standard	70	150	220
Total	108	192	300

SOLUTION Using the following definitions:

A = purchased a television with a faster refresh rate

A' = purchased a television with a standard refresh rate

B = purchased a streaming media box

B' = did not purchase a streaming media box

$$\begin{aligned} P(\text{Faster refresh rate}) &= \frac{\text{Number of faster refresh rate televisions purchased}}{\text{Total number of televisions}} \\ &= \frac{80}{300} = 0.267 \end{aligned}$$

There is a 26.7% chance that a randomly selected large-screen HDTV purchased has a faster refresh rate.

Joint Probability

Whereas simple probability refers to the probability of occurrence of simple events, **joint probability** refers to the probability of an occurrence involving two or more events. An example of joint probability is the probability that you will get heads on the first toss of a coin and heads on the second toss of a coin.

In Table 4.1 on page 154, the group of individuals who planned to purchase and actually purchased a large-screen HDTV consist only of the outcomes in the single cell “yes—planned to purchase *and* yes—actually purchased.” Because this group consists of 200 households, the probability of picking a household that planned to purchase *and* actually purchased a large-screen HDTV is

$$\begin{aligned} P(\text{Planned to purchase and actually purchased}) &= \frac{\text{Planned to purchase and actually purchased}}{\text{Total number of respondents}} \\ &= \frac{200}{1,000} = 0.20 \end{aligned}$$

Example 4.4 also demonstrates how to determine joint probability.

EXAMPLE 4.4

Determining the Joint Probability That a Household Purchased a Large-Screen HDTV with a Faster Refresh Rate and Purchased a Streaming Media Box

In Table 4.2 on page 156, the purchases are cross-classified as having a faster refresh rate or having a standard refresh rate and whether the household purchased a streaming media box. Find the probability that a randomly selected household that purchased a large-screen HDTV also purchased a television that had a faster refresh rate and purchased a streaming media box.

SOLUTION Using Equation (4.1) on page 152,

$$\begin{aligned} P(\text{Television with a faster refresh rate and streaming media box}) &= \frac{\text{Number that purchased a television with a faster refresh rate and purchased a streaming media box}}{\text{Total number of large-screen HDTV purchasers}} \\ &= \frac{38}{300} = 0.127 \end{aligned}$$

Therefore, there is a 12.7% chance that a randomly selected household that purchased a large-screen HDTV purchased a television that had a faster refresh rate and purchased a streaming media box.

Marginal Probability

The **marginal probability** of an event consists of a set of joint probabilities. You can determine the marginal probability of a particular event by using the concept of joint probability just discussed. For example, if B consists of two events, B_1 and B_2 , then $P(A)$, the probability of event A , consists of the joint probability of event A occurring with event B_1 and the joint probability of event A occurring with event B_2 . You use Equation (4.2) to compute marginal probabilities.

MARGINAL PROBABILITY

$$P(A) = P(A \text{ and } B_1) + P(A \text{ and } B_2) + \cdots + P(A \text{ and } B_k) \quad (4.2)$$

where B_1, B_2, \dots, B_k are k mutually exclusive and collectively exhaustive events, defined as follows:

Two events are **mutually exclusive** if both the events cannot occur simultaneously. A set of events is **collectively exhaustive** if one of the events must occur.

Heads and tails in a coin toss are mutually exclusive events. The result of a coin toss cannot simultaneously be a head and a tail. Heads and tails in a coin toss are also collectively exhaustive events. One of them must occur. If heads does not occur, tails must occur. If tails does not occur, heads must occur. Being male and being female are mutually exclusive and collectively exhaustive events. No person is both (the two are mutually exclusive), and everyone is one or the other (the two are collectively exhaustive).

You can use Equation (4.2) to compute the marginal probability of “planned to purchase” a large-screen HDTV:

$$\begin{aligned} P(\text{Planned to purchase}) &= P(\text{Planned to purchase and purchased}) \\ &\quad + P(\text{Planned to purchase and did not purchase}) \\ &= \frac{200}{1,000} + \frac{50}{1,000} \\ &= \frac{250}{1,000} = 0.25 \end{aligned}$$

You get the same result if you add the number of outcomes that make up the simple event “planned to purchase.”

Student Tip

The key word when using the addition rule is *or*.

General Addition Rule

How do you find the probability of event “A *or* B”? You need to consider the occurrence of either event A or event B or both A and B. For example, how can you determine the probability that a household planned to purchase *or* actually purchased a large-screen HDTV?

The event “planned to purchase *or* actually purchased” includes all households that planned to purchase and all households that actually purchased a large-screen HDTV. You examine each cell of the contingency table (Table 4.1 on page 154) to determine whether it is part of this event. From Table 4.1, the cell “planned to purchase *and* did not actually purchase” is part of the event because it includes respondents who planned to purchase. The cell “did not plan to purchase *and* actually purchased” is included because it contains respondents who actually purchased. Finally, the cell “planned to purchase *and* actually purchased” has both characteristics of interest. Therefore, one way to calculate the probability of “planned to purchase *or* actually purchased” is

$$\begin{aligned} P(\text{Planned to purchase or actually purchased}) &= P(\text{Planned to purchase and did not actually purchase}) + P(\text{Did not plan to purchase and actually purchased}) + \\ &\quad P(\text{Planned to purchase and actually purchased}) \\ &= \frac{50}{1,000} + \frac{100}{1,000} + \frac{200}{1,000} \\ &= \frac{350}{1,000} = 0.35 \end{aligned}$$

Often, it is easier to determine $P(A \text{ or } B)$, the probability of the event A *or* B, by using the **general addition rule**, defined in Equation (4.3).

GENERAL ADDITION RULE

The probability of A *or* B is equal to the probability of A plus the probability of B minus the probability of A *and* B.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{4.3}$$

Applying Equation (4.3) to the previous example produces the following result:

$$\begin{aligned}
 P(\text{Planned to purchase or actually purchased}) &= P(\text{Planned to purchase}) \\
 &\quad + P(\text{Actually purchased}) - P(\text{Planned to purchase and actually purchased}) \\
 &= \frac{250}{1,000} + \frac{300}{1,000} - \frac{200}{1,000} \\
 &= \frac{350}{1,000} = 0.35
 \end{aligned}$$

The general addition rule consists of taking the probability of A and adding it to the probability of B and then subtracting the probability of the joint event A and B from this total because the joint event has already been included in computing both the probability of A and the probability of B . Referring to Table 4.1 on page 154, if the outcomes of the event “planned to purchase” are added to those of the event “actually purchased,” the joint event “planned to purchase and actually purchased” has been included in each of these simple events. Therefore, because this joint event has been included twice, you must subtract it to compute the correct result. Example 4.5 illustrates another application of the general addition rule.

EXAMPLE 4.5

Using the General Addition Rule for the Households That Purchased Large-Screen HDTVs

In Example 4.3 on page 156, the purchases were cross-classified in Table 4.2 as televisions that had a faster refresh rate or televisions that had a standard refresh rate and whether the household purchased a streaming media box. Find the probability that among households that purchased a large-screen HDTV, they purchased a television that had a faster refresh rate or purchased a streaming media box.

SOLUTION Using Equation (4.3),

$$\begin{aligned}
 P(\text{Television had a faster refresh rate or purchased a streaming media box}) &= P(\text{Television had a faster refresh rate}) \\
 &\quad + P(\text{purchased a streaming media box}) \\
 &\quad - P(\text{Television had a faster refresh rate and purchased a streaming media box}) \\
 &= \frac{80}{300} + \frac{108}{300} - \frac{38}{300} \\
 &= \frac{150}{300} = 0.50
 \end{aligned}$$

Therefore, of households that purchased a large-screen HDTV, there is a 50% chance that a randomly selected household purchased a television that had a faster refresh rate or purchased a streaming media box.

Problems for Section 4.1

LEARNING THE BASICS

4.1 Two coins are tossed.

- a. Give an example of a simple event.
- b. Give an example of a joint event.
- c. What is the complement of a head on the first toss?
- d. What does the sample space consist of?

4.2 An urn contains 12 red balls and 8 white balls. One ball is to be selected from the urn.

- a. Give an example of a simple event.
- b. What is the complement of a red ball?
- c. What does the sample space consist of?

- 4.3** Consider the following contingency table:

	B	B'
A	10	20
A'	20	40

What is the probability of event

- a. A?
- b. A'?
- c. A and B?
- d. A or B?

- 4.4** Consider the following contingency table:

	B	B'
A	10	30
A'	25	35

What is the probability of event

- a. A'?
- b. A and B?
- c. A' and B'?
- d. A' or B'?

APPLYING THE CONCEPTS

- 4.5** For each of the following, indicate whether the type of probability involved is an example of *a priori* probability, empirical probability, or subjective probability.

- a. The next toss of a fair coin will land on heads.
- b. Italy will win soccer's World Cup the next time the competition is held.
- c. The sum of the faces of two dice will be seven.
- d. The train taking a commuter to work will be more than 10 minutes late.

- 4.6** For each of the following, state whether the events created are mutually exclusive and whether they are collectively exhaustive.

- a. Undergraduate business students were asked whether they were sophomores or juniors.
- b. Each respondent was classified by the type of car he or she drives: sedan, SUV, American, European, Asian, or none.
- c. People were asked, "Do you currently live in (i) an apartment or (ii) a house?"
- d. A product was classified as defective or not defective.

- 4.7** Which of the following events occur with a probability of zero? For each, state why or why not.

- a. A company is listed on the New York Stock Exchange and NASDAQ.
- b. A consumer owns a smartphone and a tablet.
- c. A cellphone is a Motorola and a Samsung.
- d. An automobile is a Toyota and was manufactured in the United States.

- 4.8** Do males or females feel more tense or stressed out at work? A survey of employed adults conducted online by Harris Interactive

on behalf of the American Psychological Association revealed the following:

FELT TENSE OR STRESSED OUT AT WORK		
GENDER	Yes	No
Male	244	495
Female	282	480

Source: Data extracted from "The 2013 Work and Well-Being Survey," American Psychological Association and Harris Interactive, March 2013, p. 5, bit.ly/11JGcPf.

- a. Give an example of a simple event.
- b. Give an example of a joint event.
- c. What is the complement of "Felt tense or stressed out at work"?
- d. Why is "Male and felt tense or stressed out at work" a joint event?

- 4.9** Referring to the contingency table in Problem 4.8, if an employed adult is selected at random, what is the probability that

- a. the employed adult felt tense or stressed out at work?
- b. the employed adult was a male who felt tense or stressed out at work?
- c. the employed adult was a male *or* felt tense or stressed out at work?
- d. Explain the difference in the results in (b) and (c).

- 4.10** How will marketers change their social media use in the near future? A survey by Social Media Examiner reported that 76% of B2B marketers (marketers that focus primarily on attracting businesses) plan to increase their use of LinkedIn, as compared to 55% of B2C marketers (marketers that primarily target consumers). The survey was based on 1,945 B2B marketers and 1,868 B2C marketers. The following table summarizes the results:

INCREASE USE OF LINKEDIN?	BUSINESS FOCUS		
	B2B	B2C	Total
Yes	1,478	1,027	2,505
No	467	841	1,308
Total	1,945	1,868	3,813

Source: Data extracted from "2012 Social Media Marketing Industry Report," April 2012, p. 27, bit.ly/HaWwDu.

- a. Give an example of a simple event.
- b. Give an example of a joint event.
- c. What is the complement of a marketer who plans to increase use of LinkedIn?
- d. Why is a marketer who plans to increase use of LinkedIn and is a B2C marketer a joint event?

- 4.11** Referring to the contingency table in Problem 4.10, if a marketer is selected at random, what is the probability that

- a. he or she plans to increase use of LinkedIn?
- b. he or she is a B2C marketer?
- c. he or she plans to increase use of LinkedIn *or* is a B2C marketer?
- d. Explain the difference in the results in (b) and (c).



4.12 What business and technical skills are critical for today's business intelligence/analytics and information management professionals? As part of InformationWeek's 2013 U.S. IT Salary Survey, business intelligence/analytics and information management professionals, both staff and managers, were asked to indicate what business and technical skills are critical to their job. The list of business and technical skills included *Analyzing Data*. The following table summarizes the responses to this skill:

PROFESSIONAL POSITION			
ANALYZING DATA	Staff	Management	Total
Critical	4,374	3,633	8,007
Not critical	3,436	2,631	6,067
Total	7,810	6,264	14,074

Source: Data extracted from "IT Salaries Show Slow Growth," *InformationWeek Reports*, April 2013, p. 40, ubm.io/lewjKT5.

If a professional is selected at random, what is the probability that he or she

- a. indicates analyzing data as critical to his or her job?
- b. is a manager?
- c. indicates analyzing data as critical to his or her job *or* is a manager?
- d. Explain the difference in the results in (b) and (c).

4.13 Do Americans prefer Coke or Pepsi? A survey was conducted by Public Policy Polling (PPP) in 2013; the results were as follows:

GENDER			
PREFERENCE	Female	Male	Total
Coke	120	95	215
Pepsi	95	80	175
Neither/Unsure	65	45	110
Total	280	220	500

Source: Data extracted from "Public Policy Polling" Report 2013, bit.ly/YKXfzN.

If an American is selected at random, what is the probability that he or she

- a. prefers Pepsi?
- b. is male *and* prefers Pepsi?
- c. is male *or* prefers Pepsi?
- d. Explain the difference in the results in (b) and (c).

4.14 A survey of 1,085 adults asked, "Do you enjoy shopping for clothing for yourself?" The results (data extracted from "Split Decision on Clothes Shopping," *USA Today*, January 28, 2011, p. 1B) indicated that 51% of the females enjoyed shopping for clothing for themselves as compared to 44% of the males. The sample sizes of males and females were not provided. Suppose that the results indicated that of 542 males, 238 answered yes. Of 543 females, 276 answered yes. Construct a contingency table to evaluate the probabilities. What is the probability that a respondent chosen at random

- a. enjoys shopping for clothing for himself or herself?
- b. is a female *and* enjoys shopping for clothing for herself?
- c. is a female *or* is a person who enjoys shopping for clothing?
- d. is a male *or* a female?

4.15 Each year, ratings are compiled concerning the performance of new cars during the first 90 days of use. Suppose that the cars have been categorized according to whether a car needs warranty-related repair (yes or no) and the country in which the company manufacturing a car is based (United States or not United States). Based on the data collected, the probability that the new car needs a warranty repair is 0.04, the probability that the car was manufactured by a U.S.-based company is 0.60, and the probability that the new car needs a warranty repair *and* was manufactured by a U.S.-based company is 0.025. Construct a contingency table to evaluate the probabilities of a warranty-related repair. What is the probability that a new car selected at random

- a. needs a warranty repair?
- b. needs a warranty repair *and* was manufactured by a U.S.-based company?
- c. needs a warranty repair *or* was manufactured by a U.S.-based company?
- d. needs a warranty repair *or* was not manufactured by a U.S.-based company?

4.2 Conditional Probability

Each example in Section 4.1 involves finding the probability of an event when sampling from the entire sample space. How do you determine the probability of an event if you know certain information about the events involved?

Computing Conditional Probabilities

Conditional probability refers to the probability of event *A*, given information about the occurrence of another event, *B*.

CONDITIONAL PROBABILITY

The probability of A given B is equal to the probability of A and B divided by the probability of B .

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \quad (4.4a)$$

The probability of B given A is equal to the probability of A and B divided by the probability of A .

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \quad (4.4b)$$

where

$P(A \text{ and } B)$ = joint probability of A and B

$P(A)$ = marginal probability of A

$P(B)$ = marginal probability of B

Student Tip

The variable that is *given* goes in the denominator of Equation (4.4). Since you were given planned to purchase, planned to purchase is in the denominator.

Referring to the Using Statistics scenario involving the purchase of large-screen HDTVs, suppose you were told that a household planned to purchase a large-screen HDTV. Now, what is the probability that the household actually purchased the television?

In this example, the objective is to find $P(\text{Actually purchased} | \text{Planned to purchase})$. Here you are given the information that the household planned to purchase the large-screen HDTV. Therefore, the sample space does not consist of all 1,000 households in the survey. It consists of only those households that planned to purchase the large-screen HDTV. Of 250 such households, 200 actually purchased the large-screen HDTV. Therefore, based on Table 4.1 on page 154, the probability that a household actually purchased the large-screen HDTV given that they planned to purchase is

$$\begin{aligned} P(\text{Actually purchased} | \text{Planned to purchase}) &= \frac{\text{Planned to purchase and actually purchased}}{\text{Planned to purchase}} \\ &= \frac{200}{250} = 0.80 \end{aligned}$$

You can also use Equation (4.4b) to compute this result:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

where

A = planned to purchase

B = actually purchased

then

$$\begin{aligned} P(\text{Actually purchased} | \text{Planned to purchase}) &= \frac{200/1,000}{250/1,000} \\ &= \frac{200}{250} = 0.80 \end{aligned}$$

Example 4.6 further illustrates conditional probability.

EXAMPLE 4.6

Finding the Conditional Probability of Purchasing a Streaming Media Box

Table 4.2 on page 156 is a contingency table for whether a household purchased a television with a faster refresh rate and whether the household purchased a streaming media box. If a household purchased a television with a faster refresh rate, what is the probability that it also purchased a streaming media box?

SOLUTION Because you know that the household purchased a television with a faster refresh rate, the sample space is reduced to 80 households. Of these 80 households, 38 also purchased a streaming media box. Therefore, the probability that a household purchased a streaming media box, given that the household purchased a television with a faster refresh rate, is

$$P(\text{Purchased streaming media box} \mid \text{Purchased television with faster refresh rate}) = \frac{\text{Number purchasing television with faster refresh rate and streaming media box}}{\text{Number purchasing television with faster refresh rate}} = \frac{38}{80} = 0.475$$

If you use Equation (4.4b) on page 162:

A = purchased a television with a faster refresh rate

B = purchased a streaming media box

then

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{38/300}{80/300} = 0.475$$

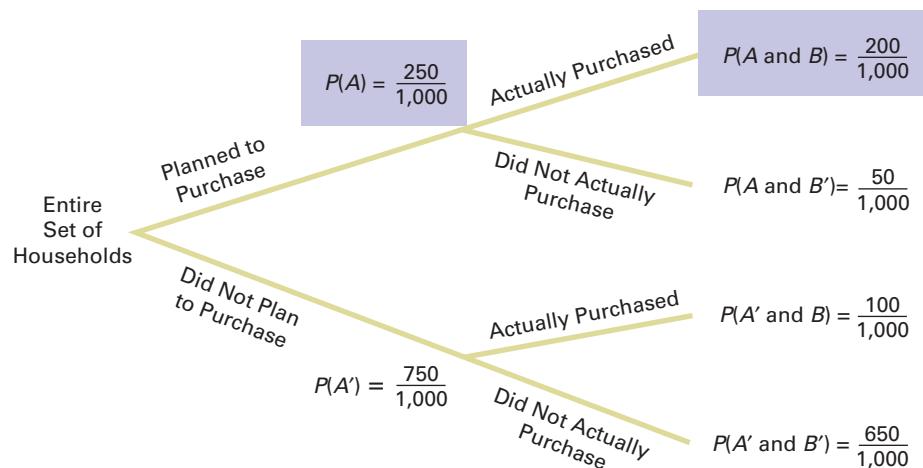
Therefore, given that the household purchased a television with a faster refresh rate, there is a 47.5% chance that the household also purchased a streaming media box. You can compare this conditional probability to the marginal probability of purchasing a streaming media box, which is $108/300 = 0.36$, or 36%. These results tell you that households that purchased televisions with a faster refresh rate are more likely to purchase a streaming media box than are households that purchased large-screen HDTVs that have a standard refresh rate.

Decision Trees

In Table 4.1 on page 154, households are classified according to whether they planned to purchase and whether they actually purchased large-screen HDTVs. A **decision tree** is an alternative to the contingency table. Figure 4.3 represents the decision tree for this example.

FIGURE 4.3

Decision tree for planned to purchase and actually purchased



In Figure 4.3, beginning at the left with the entire set of households, there are two “branches” for whether or not the household planned to purchase a large-screen HDTV. Each of these branches has two subbranches, corresponding to whether the household actually purchased or did not actually purchase the large-screen HDTV. The probabilities at the end of the initial branches represent the marginal probabilities of A and A' . The probabilities at the end of each of the four subbranches represent the joint probability for each combination of events A and B . You compute the conditional probability by dividing the joint probability by the appropriate marginal probability.

For example, to compute the probability that the household actually purchased, given that the household planned to purchase the large-screen HDTV, you take $P(\text{Planned to purchase and actually purchased})$ and divide by $P(\text{Planned to purchase})$. From Figure 4.3,

$$\begin{aligned} P(\text{Actually purchased} \mid \text{Planned to purchase}) &= \frac{200/1,000}{250/1,000} \\ &= \frac{200}{250} = 0.80 \end{aligned}$$

Example 4.7 illustrates how to construct a decision tree.

EXAMPLE 4.7

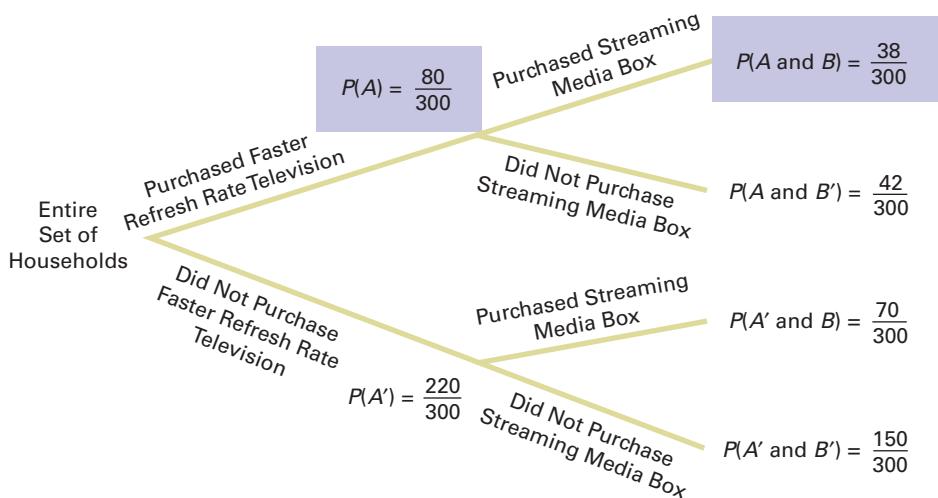
Constructing the Decision Tree for the Households That Purchased Large-Screen HDTVs

FIGURE 4.4

Decision tree for purchased a television with a faster refresh rate and a streaming media box

Using the cross-classified data in Table 4.2 on page 156, construct the decision tree. Use the decision tree to find the probability that a household purchased a streaming media box, given that the household purchased a television with a faster refresh rate.

SOLUTION The decision tree for purchased a streaming media box and a television with a faster refresh rate is displayed in Figure 4.4.



Using Equation (4.4b) on page 162 and the following definitions,

A = purchased a television with a faster refresh rate

B = purchased a streaming media box

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{38/300}{80/300} = 0.475$$

Independence

In the example concerning the purchase of large-screen HDTVs, the conditional probability is $200/250 = 0.80$ that the selected household actually purchased the large-screen HDTV, given that the household planned to purchase. The simple probability of selecting a household that actually purchased is $300/1,000 = 0.30$. This result shows that the prior knowledge that the household planned to purchase affected the probability that the household actually purchased the television. In other words, the outcome of one event is *dependent* on the outcome of a second event.

When the outcome of one event does *not* affect the probability of occurrence of another event, the events are said to be independent. **Independence** can be determined by using Equation (4.5).

INDEPENDENCE

Two events, A and B , are independent if and only if

$$P(A | B) = P(A) \quad (4.5)$$

where

$P(A | B)$ = conditional probability of A given B

$P(A)$ = marginal probability of A

Example 4.8 demonstrates the use of Equation (4.5).

EXAMPLE 4.8

Determining Independence

TABLE 4.3

Satisfaction with Purchase of Large-Screen HDTVs

In the follow-up survey of the 300 households that actually purchased large-screen HDTVs, the households were asked if they were satisfied with their purchases. Table 4.3 cross-classifies the responses to the satisfaction question with the responses to whether the television had a faster refresh rate.

TELEVISION REFRESH RATE	SATISFIED WITH PURCHASE?		
	Yes	No	Total
Faster	64	16	80
Standard	176	44	220
Total	240	60	300

Determine whether being satisfied with the purchase and the refresh rate of the television purchased are independent.

SOLUTION For these data,

$$P(\text{Satisfied} | \text{Faster refresh rate}) = \frac{64/300}{80/300} = \frac{64}{80} = 0.80$$

which is equal to

$$P(\text{Satisfied}) = \frac{240}{300} = 0.80$$

Thus, being satisfied with the purchase and the refresh rate of the television purchased are independent. Knowledge of one event does not affect the probability of the other event.

Multiplication Rules

The **general multiplication rule** is derived using Equation (4.4a) on page 162:

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$

and solving for the joint probability $P(A \text{ and } B)$.

GENERAL MULTIPLICATION RULE

The probability of A and B is equal to the probability of A given B times the probability of B .

$$P(A \text{ and } B) = P(A | B)P(B) \quad (4.6)$$

Example 4.9 demonstrates the use of the general multiplication rule.

EXAMPLE 4.9

Using the General Multiplication Rule

Consider the 80 households that purchased televisions that had a faster refresh rate. In Table 4.3 on page 165, you see that 64 households are satisfied with their purchase, and 16 households are dissatisfied. Suppose 2 households are randomly selected from the 80 households. Find the probability that both households are satisfied with their purchase.

SOLUTION Here you can use the multiplication rule in the following way. If

$$\begin{aligned} A &= \text{second household selected is satisfied} \\ B &= \text{first household selected is satisfied} \end{aligned}$$

then, using Equation (4.6),

$$P(A \text{ and } B) = P(A | B)P(B)$$

The probability that the first household is satisfied with the purchase is $64/80$. However, the probability that the second household is also satisfied with the purchase depends on the result of the first selection. If the first household is not returned to the sample after the satisfaction level is determined (i.e., sampling without replacement), the number of households remaining is 79. If the first household is satisfied, the probability that the second is also satisfied is $63/79$ because 63 satisfied households remain in the sample. Therefore,

$$P(A \text{ and } B) = \left(\frac{63}{79}\right)\left(\frac{64}{80}\right) = 0.6380$$

There is a 63.80% chance that both of the households sampled will be satisfied with their purchase.

The **multiplication rule for independent events** is derived by substituting $P(A)$ for $P(A | B)$ in Equation (4.6).

MULTIPLICATION RULE FOR INDEPENDENT EVENTS

If A and B are independent, the probability of A and B is equal to the probability of A times the probability of B .

$$P(A \text{ and } B) = P(A)P(B) \quad (4.7)$$

If this rule holds for two events, A and B , then A and B are independent. Therefore, there are two ways to determine independence:

1. Events A and B are independent if, and only if, $P(A|B) = P(A)$.
2. Events A and B are independent if, and only if, $P(A \text{ and } B) = P(A)P(B)$.

Marginal Probability Using the General Multiplication Rule

In Section 4.1, marginal probability was defined using Equation (4.2) on page 157. You can state the equation for marginal probability by using the general multiplication rule. If

$$P(A) = P(A \text{ and } B_1) + P(A \text{ and } B_2) + \cdots + P(A \text{ and } B_k)$$

then, using the general multiplication rule, Equation (4.8) defines the marginal probability.

MARGINAL PROBABILITY USING THE GENERAL MULTIPLICATION RULE

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k) \quad (4.8)$$

where B_1, B_2, \dots, B_k are k mutually exclusive and collectively exhaustive events.

To illustrate Equation (4.8), refer to Table 4.1 on page 154. Let

$P(A)$ = probability of planned to purchase

$P(B_1)$ = probability of actually purchased

$P(B_2)$ = probability of did not actually purchase

Then, using Equation (4.8), the probability of planned to purchase is

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) \\ &= \left(\frac{200}{300}\right)\left(\frac{300}{1,000}\right) + \left(\frac{50}{700}\right)\left(\frac{700}{1,000}\right) \\ &= \frac{200}{1,000} + \frac{50}{1,000} = \frac{250}{1,000} = 0.25 \end{aligned}$$

Problems for Section 4.2

LEARNING THE BASICS

4.16 Consider the following contingency table:

	B	B'
A	10	20
A'	20	40

What is the probability of

- a. $A|B$?
- b. $A'|B$?
- c. $A'|B'$?
- d. Are events A and B independent?

4.17 Consider the following contingency table:

	B	B'
A	10	30
A'	25	35

What is the probability of

- a. $A|B$?
- b. $A'|B$?
- c. $A|B'$?
- d. Are events A and B independent?

4.18 If $P(A \text{ and } B) = 0.4$ and $P(B) = 0.8$, find $P(A|B)$.

4.19 If $P(A) = 0.7$, $P(B) = 0.6$, and A and B are independent, find $P(A \text{ and } B)$.

4.20 If $P(A) = 0.3$, $P(B) = 0.4$, and $P(A \text{ and } B) = 0.2$, are A and B independent?

APPLYING THE CONCEPTS

4.21 Do males or females feel more tense or stressed out at work? A survey of employed adults conducted online by Harris Interactive on behalf of the American Psychological Association revealed the following:

FELT TENSE OR STRESSED OUT AT WORK		
GENDER	Yes	No
Male	244	495
Female	282	480

Source: Data extracted from "The 2013 Work and Well-Being Survey," American Psychological Association and Harris Interactive, March 2013, p. 5, bit.ly/11JGcPf.

- a. Given that the employed adult felt tense or stressed out at work, what is the probability that the employed adult was a male?
- b. Given that the employed adult is male, what is the probability that he felt tense or stressed out at work?
- c. Explain the difference in the results in (a) and (b).
- d. Is feeling tense or stressed out at work and gender independent?

4.22 How will marketers change their social media use in the near future? A survey by Social Media Examiner reported that 76% of B2B marketers (marketers that focus primarily on attracting businesses) plan to increase their use of LinkedIn, as compared to 55% of B2C marketers (marketers that primarily target consumers). The survey was based on 1,945 B2B marketers and 1,868 B2C marketers. The following table summarizes the results:

INCREASE USE OF LINKEDIN?	BUSINESS FOCUS		
	B2B	B2C	Total
Yes	1,478	1,027	2,505
No	467	841	1,308
Total	1,945	1,868	3,813

Source: Data extracted from "2012 Social Media Marketing Industry Report," April 2012, p. 27, bit.ly/HaWwDu.

- a. Suppose you know that the marketer is a B2B marketer. What is the probability that he or she plans to increase use of LinkedIn?
- b. Suppose you know that the marketer is a B2C marketer. What is the probability that he or she plans to increase use of LinkedIn?
- c. Are the two events, increase use of LinkedIn and business focus, independent? Explain.

4.23 Do Americans prefer Coke or Pepsi? A survey was conducted by Public Policy Polling (PPP) in 2013; the results were as follows:

PREFERENCE	GENDER		
	Female	Male	Total
Coke	120	95	215
Pepsi	95	80	175
Neither/Unsure	65	45	110
Total	280	220	500

Source: Data extracted from "Public Policy Polling" Report 2013, bit.ly/YKXfzN.

- a. Given that an American is a male, what is the probability that he prefers Pepsi?
- b. Given that an American is a female, what is the probability that she prefers Pepsi?
- c. Is preference independent of gender? Explain.

4.24 What business and technical skills are critical for today's business intelligence/analytics and information management professionals? As part of InformationWeek's 2013 U.S. IT Salary Survey, business intelligence/analytics and information management professionals, both staff and managers, were asked to indicate what business and technical skills are critical to their job. The list of business and technical skills included *Analyzing Data*. The following table summarizes the responses to this skill:

ANALYZING DATA	PROFESSIONAL POSITION		
	Staff	Management	Total
Critical	4,374	3,633	8,007
Not critical	3,436	2,631	6,067
Total	7,810	6,264	14,074

Source: Data extracted from "IT Salaries Show Slow Growth," *InformationWeek Reports*, April 2013, p. 40, ubm.io/1ewjKT5.

- a. Given that a professional is staff, what is the probability that the professional indicates analyzing data as critical to his or her job?
- b. Given that a professional is staff, what is the probability that the professional does not indicate analyzing data as critical to his or her job?
- c. Given that a professional is a manager, what is the probability that the professional indicates analyzing data as critical to his or her job?
- d. Given that a professional is a manager, what is the probability that the professional does not indicate analyzing data as critical to his or her job?

4.25 A survey of 1,085 adults asked, “Do you enjoy shopping for clothing for yourself?” The results (data extracted from “Split Decision on Clothes Shopping,” *USA Today*, January 28, 2011, p. 1B) indicated that 51% of the females enjoyed shopping for clothing for themselves as compared to 44% of the males. The sample sizes of males and females were not provided. Suppose that the results were as shown in the following table:

ENJOYS SHOPPING FOR CLOTHING	GENDER		
	Male	Female	Total
Yes	238	276	514
No	304	267	571
Total	542	543	1,085

- a. Suppose that the respondent chosen is a female. What is the probability that she does not enjoy shopping for clothing?
- b. Suppose that the respondent chosen enjoys shopping for clothing. What is the probability that the individual is a male?
- c. Are enjoying shopping for clothing and the gender of the individual independent? Explain.

4.26 Each year, ratings are compiled concerning the performance of new cars during the first 90 days of use. Suppose that the cars have been categorized according to whether a car needs warranty-related repair (yes or no) and the country in which the company manufacturing a car is based (United States or not United States). Based on the data collected, the probability that the new car needs a warranty repair is 0.04, the probability that the car is manufactured by a U.S.-based company is 0.60, and the probability that the new car needs a warranty repair *and* was manufactured by a U.S.-based company is 0.025.

- a. Suppose you know that a company based in the United States manufactured a particular car. What is the probability that the car needs a warranty repair?
- b. Suppose you know that a company based in the United States did not manufacture a particular car. What is the probability that the car needs a warranty repair?
- c. Are need for a warranty repair and location of the company manufacturing the car independent?

4.27 In 40 of the 62 years from 1950 through 2012 (in 2011 there was virtually no change), the S&P 500 finished higher after the first five days of trading. In 35 of those 40 years, the S&P 500 finished higher for the year. Is a good first week a good omen for the upcoming year? The following table gives the first-week and annual performance over this 62-year period:

S&P 500'S ANNUAL PERFORMANCE		
FIRST WEEK	Higher	Lower
Higher	35	5
Lower	11	11

- a. If a year is selected at random, what is the probability that the S&P 500 finished higher for the year?
- b. Given that the S&P 500 finished higher after the first five days of trading, what is the probability that it finished higher for the year?
- c. Are the two events “first-week performance” and “annual performance” independent? Explain.
- d. Look up the performance after the first five days of 2013 and the 2013 annual performance of the S&P 500 at finance.yahoo.com. Comment on the results.

4.28 A standard deck of cards is being used to play a game. There are four suits (hearts, diamonds, clubs, and spades), each having 13 faces (ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, jack, queen, and king), making a total of 52 cards. This complete deck is thoroughly mixed, and you will receive the first 2 cards from the deck, without replacement (the first card is not returned to the deck after it is selected).

- a. What is the probability that both cards are queens?
- b. What is the probability that the first card is a 10 and the second card is a 5 or 6?
- c. If you were sampling with replacement (the first card is returned to the deck after it is selected), what would be the answer in (a)?
- d. In the game of blackjack, the face cards (jack, queen, king) count as 10 points, and the ace counts as either 1 or 11 points. All other cards are counted at their face value. Blackjack is achieved if 2 cards total 21 points. What is the probability of getting blackjack in this problem?

4.29 A box of nine gloves contains two left-handed gloves and seven right-handed gloves.

- a. If two gloves are randomly selected from the box, without replacement (the first glove is not returned to the box after it is selected), what is the probability that both gloves selected will be right-handed?
- b. If two gloves are randomly selected from the box, without replacement (the first glove is not returned to the box after it is selected), what is the probability that there will be one right-handed glove and one left-handed glove selected?
- c. If three gloves are selected, with replacement (the gloves are returned to the box after they are selected), what is the probability that all three will be left-handed?
- d. If you were sampling with replacement (the first glove is returned to the box after it is selected), what would be the answers to (a) and (b)?

4.3 Bayes' Theorem

Bayes' theorem is used to revise previously calculated probabilities based on new information. Developed by Thomas Bayes in the eighteenth century (see references 1 and 6), Bayes' theorem is an extension of what you previously learned about conditional probability.

You can apply Bayes' theorem to the situation in which M&R Electronics World is considering marketing a new model of televisions. In the past, 40% of the new-model televisions have been successful, and 60% have been unsuccessful. Before introducing the new-model

television, the marketing research department conducts an extensive study and releases a report, either favorable or unfavorable. In the past, 80% of the successful new-model television(s) had received favorable market research reports, and 30% of the unsuccessful new-model television(s) had received favorable reports. For the new model of television under consideration, the marketing research department has issued a favorable report. What is the probability that the television will be successful?

Bayes' theorem is developed from the definition of conditional probability. To find the conditional probability of B given A , consider Equation (4.4b) (originally presented on page 162 and shown again below):

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

Bayes' theorem is derived by substituting Equation (4.8) on page 167 for $P(A)$ in the denominator of Equation (4.4b).

BAYES' THEOREM

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_k)P(B_k)} \quad (4.9)$$

where B_i is the i th event out of k mutually exclusive and collectively exhaustive events.

To use Equation (4.9) for the television-marketing example, let

$$\begin{array}{ll} \text{event } S = \text{successful television} & \text{event } F = \text{favorable report} \\ \text{event } S' = \text{unsuccessful television} & \text{event } F' = \text{unfavorable report} \end{array}$$

and

$$\begin{array}{ll} P(S) = 0.40 & P(F|S) = 0.80 \\ P(S') = 0.60 & P(F|S') = 0.30 \end{array}$$

Then, using Equation (4.9),

$$\begin{aligned} P(S|F) &= \frac{P(F|S)P(S)}{P(F|S)P(S) + P(F|S')P(S')} \\ &= \frac{(0.80)(0.40)}{(0.80)(0.40) + (0.30)(0.60)} \\ &= \frac{0.32}{0.32 + 0.18} = \frac{0.32}{0.50} \\ &= 0.64 \end{aligned}$$

The probability of a successful television, given that a favorable report was received, is 0.64. Thus, the probability of an unsuccessful television, given that a favorable report was received, is $1 - 0.64 = 0.36$.

Table 4.4 summarizes the computation of the probabilities, and Figure 4.5 presents the decision tree.

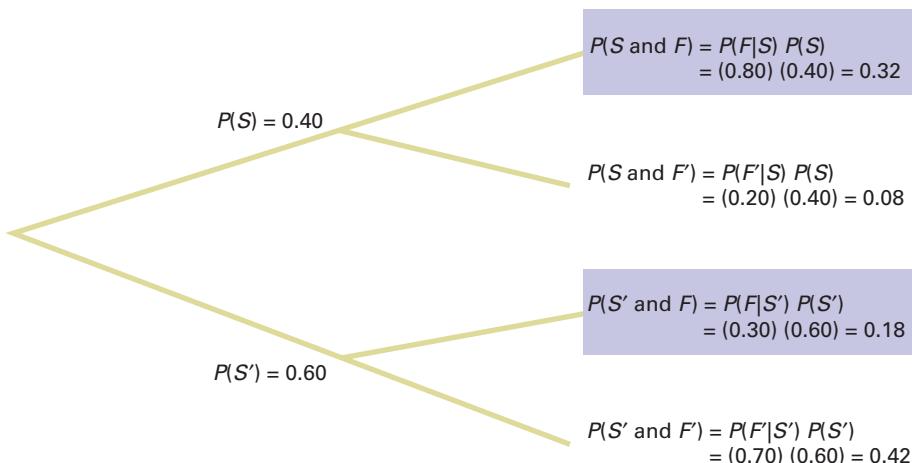
TABLE 4.4

Bayes' Theorem Computations for the Television-Marketing Example

Event S_i	Prior Probability $P(S_i)$	Conditional Probability $P(F S_i)$	Joint Probability $P(F S_i) P(S_i)$	Revised Probability $P(S_i F)$
$S = \text{successful television}$	0.40	0.80	0.32	$P(S F) = 0.32/0.50 = 0.64$
$S' = \text{unsuccessful television}$	0.60	0.30	$\frac{0.18}{0.50}$	$P(S' F) = 0.18/0.50 = 0.36$

FIGURE 4.5

Decision tree for marketing a new television



Example 4.10 applies Bayes' theorem to a medical diagnosis problem.

EXAMPLE 4.10

Using Bayes' Theorem in a Medical Diagnosis Problem

The probability that a person has a certain disease is 0.03. Medical diagnostic tests are available to determine whether the person actually has the disease. If the disease is actually present, the probability that the medical diagnostic test will give a positive result (indicating that the disease is present) is 0.90. If the disease is not actually present, the probability of a positive test result (indicating that the disease is present) is 0.02. Suppose that the medical diagnostic test has given a positive result (indicating that the disease is present). What is the probability that the disease is actually present? What is the probability of a positive test result?

SOLUTION Let

$$\begin{array}{ll} \text{event } D = \text{has disease} & \text{event } T = \text{test is positive} \\ \text{event } D' = \text{does not have disease} & \text{event } T' = \text{test is negative} \end{array}$$

and

$$\begin{array}{ll} P(D) = 0.03 & P(T|D) = 0.90 \\ P(D') = 0.97 & P(T|D') = 0.02 \end{array}$$

Using Equation (4.9) on page 170,

$$\begin{aligned} P(D|T) &= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D')P(D')} \\ &= \frac{(0.90)(0.03)}{(0.90)(0.03) + (0.02)(0.97)} \\ &= \frac{0.0270}{0.0270 + 0.0194} = \frac{0.0270}{0.0464} \\ &= 0.582 \end{aligned}$$

(continued)

The probability that the disease is actually present, given that a positive result has occurred (indicating that the disease is present), is 0.582. Table 4.5 summarizes the computation of the probabilities, and Figure 4.6 presents the decision tree. The denominator in Bayes' theorem represents $P(T)$, the probability of a positive test result, which in this case is 0.0464, or 4.64%.

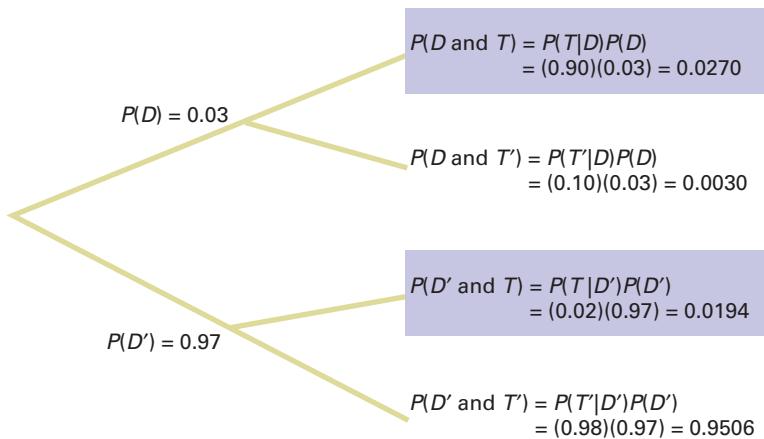
TABLE 4.5

Bayes' Theorem Computations for the Medical Diagnosis Problem

Event D_i	Prior Probability $P(D_i)$	Conditional Probability $P(T D_i)$	Joint Probability $P(T D_i)P(D_i)$	Revised Probability $P(D_i T)$
$D = \text{has disease}$	0.03	0.90	0.0270	$P(D T) = 0.0270/0.0464 = 0.582$
$D' = \text{does not have disease}$	0.97	0.02	$\frac{0.0194}{0.0464}$	$P(D' T) = 0.0194/0.0464 = 0.418$

FIGURE 4.6

Decision tree for a medical diagnosis problem



THINK ABOUT THIS

Divine Providence and Spam

Would you ever guess that the essays *Divine Benevolence: Or, An Attempt to Prove That the Principal End of the Divine Providence and Government Is the Happiness of His Creatures* and *An Essay Towards Solving a Problem in the Doctrine of Chances* were written by the same person? Probably not, and in doing so, you illustrate a modern-day application of Bayesian statistics: spam, or junk mail filters.

In not guessing correctly, you probably looked at the words in the titles of the essays and concluded that they were talking about two different things. An implicit rule you used was that word frequencies vary by subject matter. A statistics essay would very likely contain the word *statistics* as well as words such as *chance*, *problem*, and *solving*. An eighteenth-century essay about theology and religion would be more likely to contain the uppercase forms of *Divine* and *Providence*.

Likewise, there are words you would guess to be very unlikely to appear in either book, such as technical terms from finance, and words that are most likely to appear in both—common words

such as *a*, *and*, and *the*. That words would be either likely or unlikely suggests an application of probability theory. Of course, likely and unlikely are fuzzy concepts, and we might occasionally misclassify an essay if we kept things too simple, such as relying solely on the occurrence of the words *Divine* and *Providence*.

For example, a profile of the late Harris Milstead, better known as *Divine*, the star of *Hairspray* and other films, visiting Providence (Rhode Island), would most certainly not be an essay about theology. But if we widened the number of words we examined and found such words as *movie* or the name John Waters (*Divine's* director in many films), we probably would quickly realize the essay had something to do with twentieth-century cinema and little to do with theology and religion.

We can use a similar process to try to classify a new email message in your in-box as either spam or a legitimate message (called “ham,” in this context). We would first need to add to your email program a “spam filter” that has the ability to track word frequencies associated with spam and

ham messages as you identify them on a day-to-day basis. This would allow the filter to constantly update the prior probabilities necessary to use Bayes' theorem. With these probabilities, the filter can ask, “What is the probability that an email is spam, given the presence of a certain word?”

Applying the terms of Equation (4.9) on page 170, such a Bayesian spam filter would multiply the probability of finding the word in a spam email, $P(A|B)$, by the probability that the email is spam, $P(B)$, and then divide by the probability of finding the word in an email, the denominator in Equation (4.9). Bayesian spam filters also use shortcuts by focusing on a small set of words that have a high probability of being found in a spam message as well as on a small set of other words that have a low probability of being found in a spam message.

As spammers (people who send junk email) learned of such new filters, they tried to outfox them. Having learned that Bayesian filters might be assigning a high $P(A|B)$ value to words commonly found in spam, such as *Viagra*, spammers thought they could fool the filter by misspelling

the word as Vi@gr@ or V1agra. What they overlooked was that the misspelled variants were even *more likely* to be found in a spam message than the original word. Thus, the misspelled variants made the job of spotting spam *easier* for the Bayesian filters.

Other spammers tried to fool the filters by adding "good" words, words that would have a low probability of being found in a spam message, or "rare" words, words not frequently encountered in any message. But these spammers overlooked the fact that the conditional probabilities are constantly updated and that words once considered "good" would be soon discarded from the good list by the filter as their $P(A|B)$, value increased. Likewise, as "rare" words grew more common in spam and yet stayed rare in ham, such words

acted like the misspelled variants that others had tried earlier.

Even then, and perhaps after reading about Bayesian statistics, spammers thought that they could "break" Bayesian filters by inserting random words in their messages. Those random words would affect the filter by causing it to see many words whose $P(A|B)$, value would be low. The Bayesian filter would begin to label many spam messages as ham and end up being of no practical use. Spammers again overlooked that conditional probabilities are constantly updated.

Other spammers decided to eliminate all or most of the words in their messages and replace them with graphics so that Bayesian filters would have very few words with which to form conditional probabilities. But this approach failed, too, as

Bayesian filters were rewritten to consider things other than words in a message. After all, Bayes' theorem concerns *events*, and "graphics present with no text" is as valid an event as "some word, X , present in a message." Other future tricks will ultimately fail for the same reason. (By the way, spam filters use non-Bayesian techniques as well, which makes spammers' lives even more difficult.)

Bayesian spam filters are an example of the unexpected way that applications of statistics can show up in your daily life. You will discover more examples as you read the rest of this book. By the way, the author of the two essays mentioned earlier was Thomas Bayes, who is a lot more famous for the second essay than the first essay, a failed attempt to use mathematics and logic to prove the existence of God.

Problems for Section 4.3

LEARNING THE BASICS

4.30 If $P(B) = 0.05$, $P(A|B) = 0.80$, $P(B') = 0.95$, and $P(A|B') = 0.40$, find $P(B|A)$.

4.31 If $P(B) = 0.30$, $P(A|B) = 0.60$, $P(B') = 0.70$, and $P(A|B') = 0.50$, find $P(B|A)$.

APPLYING THE CONCEPTS

4.32 In Example 4.10 on page 171, suppose that the probability that a medical diagnostic test will give a positive result if the disease is not present is reduced from 0.02 to 0.01.

- If the medical diagnostic test has given a positive result (indicating that the disease is present), what is the probability that the disease is actually present?
- If the medical diagnostic test has given a negative result (indicating that the disease is not present), what is the probability that the disease is not present?

4.33 An advertising executive is studying television viewing habits of married men and women during prime-time hours. Based on past viewing records, the executive has determined that during prime time, husbands are watching television 60% of the time. When the husband is watching television, 40% of the time the wife is also watching. When the husband is not watching television, 30% of the time the wife is watching television.

- Find the probability that if the wife is watching television, the husband is also watching television.
- Find the probability that the wife is watching television during prime time.

 **4.34** Olive Construction Company is determining whether it should submit a bid for a new shopping center. In the past, Olive's main competitor, Base Construction Company, has submitted bids 70% of the time. If Base Construction Company does not bid on a job, the probability that Olive Construction Company will get the job is 0.50. If Base Construction Company bids on a job, the probability that Olive Construction Company will get the job is 0.25.

- If Olive Construction Company gets the job, what is the probability that Base Construction Company did not bid?

- What is the probability that Olive Construction Company will get the job?

4.35 Laid-off workers who become entrepreneurs because they cannot find meaningful employment with another company are known as *entrepreneurs by necessity*. *The Wall Street Journal* reported that these entrepreneurs by necessity are less likely to grow into large businesses than are *entrepreneurs by choice*. (Source: J. Bailey, "Desire—More Than Need—Builds a Business," *The Wall Street Journal*, May 21, 2001, p. B4.) This article states that 89% of the entrepreneurs in the United States are entrepreneurs by choice and 11% are entrepreneurs by necessity. Only 2% of entrepreneurs by necessity expect their new business to employ 20 or more people within five years, whereas 14% of entrepreneurs by choice expect to employ at least 20 people within five years.

- If an entrepreneur is selected at random and that individual expects that his or her new business will employ 20 or more people within five years, what is the probability that this individual is an entrepreneur by choice?
- Discuss several possible reasons why entrepreneurs by choice are more likely than entrepreneurs by necessity to believe that they will grow their businesses.

4.36 The editor of a textbook publishing company is trying to decide whether to publish a proposed business statistics textbook. Information on previous textbooks published indicates that 10% are huge successes, 20% are modest successes, 40% break even, and 30% are losers. However, before a publishing decision is made, the book will be reviewed. In the past, 99% of the huge successes received favorable reviews, 70% of the moderate successes received favorable reviews, 40% of the break-even books received favorable reviews, and 20% of the losers received favorable reviews.

- If the proposed textbook receives a favorable review, how should the editor revise the probabilities of the various outcomes to take this information into account?
- What proportion of textbooks receive favorable reviews?

4.37 A municipal bond service has three rating categories (*A*, *B*, and *C*). Suppose that in the past year, of the municipal bonds issued throughout the United States, 70% were rated *A*, 20% were rated *B*, and 10% were rated *C*. Of the municipal bonds rated *A*, 50% were issued by cities, 40% by suburbs, and 10% by rural areas. Of the municipal bonds rated *B*, 60% were issued by cities, 20% by

suburbs, and 20% by rural areas. Of the municipal bonds rated *C*, 90% were issued by cities, 5% by suburbs, and 5% by rural areas.

- If a new municipal bond is to be issued by a city, what is the probability that it will receive an *A* rating?
- What proportion of municipal bonds are issued by cities?
- What proportion of municipal bonds are issued by suburbs?

4.4 Counting Rules

In Equation (4.1) on page 152, the probability of occurrence of an outcome was defined as the number of ways the outcome occurs, divided by the total number of possible outcomes. Often, there are a large number of possible outcomes, and determining the exact number can be difficult. In such circumstances, rules have been developed for counting the number of possible outcomes. This section presents five different counting rules.

Counting Rule 1 Counting rule 1 determines the number of possible outcomes for a set of mutually exclusive and collectively exhaustive events.

COUNTING RULE 1

If any one of k different mutually exclusive and collectively exhaustive events can occur on each of n trials, the number of possible outcomes is equal to

$$k^n \quad (4.10)$$

For example, using Equation (4.10), the number of different possible outcomes from tossing a two-sided coin five times is $2^5 = 2 \times 2 \times 2 \times 2 \times 2 = 32$.

EXAMPLE 4.11

Rolling a Die Twice

Suppose you roll a die twice. How many different possible outcomes can occur?

SOLUTION If a six-sided die is rolled twice, using Equation (4.10), the number of different outcomes is $6^2 = 36$.

Counting Rule 2 The second counting rule is a more general version of the first counting rule and allows the number of possible events to differ from trial to trial.

COUNTING RULE 2

If there are k_1 events on the first trial, k_2 events on the second trial, . . . , and k_n events on the n th trial, then the number of possible outcomes is

$$(k_1)(k_2) \dots (k_n) \quad (4.11)$$

For example, a state motor vehicle department would like to know how many license plate numbers are available if a license plate number consists of three letters followed by three numbers (0 through 9). Using Equation (4.11), if a license plate number consists of three letters followed by three numbers, the total number of possible outcomes is $(26)(26)(26)(10)(10)(10) = 17,576,000$.

EXAMPLE 4.12**Determining the Number of Different Dinners**

A restaurant menu has a price-fixed complete dinner that consists of an appetizer, an entrée, a beverage, and a dessert. You have a choice of 5 appetizers, 10 entrées, 3 beverages, and 6 desserts. Determine the total number of possible dinners.

SOLUTION Using Equation (4.11), the total number of possible dinners is $(5)(10)(3)(6) = 900$.

Counting Rule 3 The third counting rule involves computing the number of ways that a set of items can be arranged in order.

COUNTING RULE 3

The number of ways that all n items can be arranged in order is

$$n! = (n)(n - 1) \dots (1) \quad (4.12)$$

where $n!$ is called n factorial, and $0!$ is defined as 1.

EXAMPLE 4.13**Using Counting Rule 3**

If a set of six books is to be placed on a shelf, in how many ways can the six books be arranged?

SOLUTION To begin, you must realize that any of the six books could occupy the first position on the shelf. Once the first position is filled, there are five books to choose from in filling the second position. You continue this assignment procedure until all the positions are occupied. The number of ways that you can arrange six books is

$$n! = 6! = (6)(5)(4)(3)(2)(1) = 720$$

Counting Rule 4 In many instances you need to know the number of ways in which a subset of an entire group of items can be arranged in *order*. Each possible arrangement is called a **permutation**.

Student Tip

Both permutations and combinations assume that you are sampling without replacement.

COUNTING RULE 4: PERMUTATIONS

The number of ways of arranging x objects selected from n objects in order is

$${}_nP_x = \frac{n!}{(n - x)!} \quad (4.13)$$

where

n = total number of objects

x = number of objects to be arranged

$n!$ = n factorial = $n(n - 1) \dots (1)$

P = symbol for permutations¹

¹On many scientific calculators, there is a button labeled nPr that allows you to compute permutations. The symbol r is used instead of x .

EXAMPLE 4.14**Using Counting Rule 4**

Modifying Example 4.13, if you have six books, but there is room for only four books on the shelf, in how many ways can you arrange these books on the shelf?

SOLUTION Using Equation (4.13), the number of ordered arrangements of four books selected from six books is equal to

$${}_nP_x = \frac{n!}{(n-x)!} = \frac{6!}{(6-4)!} = \frac{(6)(5)(4)(3)(2)(1)}{(2)(1)} = 360$$

Counting Rule 5 In many situations, you are not interested in the *order* of the outcomes but only in the number of ways that x items can be selected from n items, *irrespective of order*. Each possible selection is called a **combination**.

COUNTING RULE 5: COMBINATIONS

The number of ways of selecting x objects from n objects, irrespective of order, is equal to

$${}_nC_x = \frac{n!}{x!(n-x)!} \quad (4.14)$$

where

n = total number of objects

x = number of objects to be arranged

$n!$ = n factorial = $n(n-1)\dots(1)$

C = symbol for combinations²

²On many scientific calculators, there is a button labeled nCr that allows you to compute combinations. The symbol r is used instead of x .

If you compare this rule to counting rule 4, you see that it differs only in the inclusion of a term $x!$ in the denominator. When permutations were used, all of the arrangements of the x objects are distinguishable. With combinations, the $x!$ possible arrangements of objects are irrelevant.

EXAMPLE 4.15**Using Counting Rule 5**

Modifying Example 4.14, if the order of the books on the shelf is irrelevant, in how many ways can you arrange these books on the shelf?

SOLUTION Using Equation (4.14), the number of combinations of four books selected from six books is equal to

$${}_nC_x = \frac{n!}{x!(n-x)!} = \frac{6!}{4!(6-4)!} = \frac{(6)(5)(4)(3)(2)(1)}{(4)(3)(2)(1)(2)(1)} = 15$$

Problems for Section 4.4**APPLYING THE CONCEPTS**

-  **4.38** If there are 10 multiple-choice questions on an exam, each having three possible answers, how many different sequences of answers are there?

- 4.39** A lock on a bank vault consists of three dials, each with 30 positions. In order for the vault to open, each of the three dials must be in the correct position.

- a. How many different possible dial combinations are there for this lock?

- b. What is the probability that if you randomly select a position on each dial, you will be able to open the bank vault?
- c. Explain why “dial combinations” are not mathematical combinations expressed by Equation (4.14).

4.40 a. If a coin is tossed seven times, how many different outcomes are possible?

b. If a die is tossed seven times, how many different outcomes are possible?

c. Discuss the differences in your answers to (a) and (b).

4.41 A particular brand of women’s jeans is available in seven different sizes, three different colors, and three different styles. How many different women’s jeans does the store manager need to order to have one pair of each type?

4.42 You would like to make a salad that consists of lettuce, tomato, cucumber, and peppers. You go to the supermarket, intending to purchase one variety of each of these ingredients. You discover that there are eight varieties of lettuce, four varieties of tomatoes, three varieties of cucumbers, and three varieties of peppers for sale at the supermarket. If you buy them all, how many different salads can you make?

4.43 A team is being formed that includes four different people. There are four different positions on the teams. How many different ways are there to assign the four people to the four positions?

4.44 In Major League Baseball, there are five teams in the Eastern Division of the National League: Atlanta, Florida, New York,

Philadelphia, and Washington. How many different orders of finish are there for these five teams? (Assume that there are no ties in the standings.) Do you believe that all these orders are equally likely? Discuss.

4.45 Referring to Problem 4.44, how many different orders of finish are possible for the first four positions?

4.46 A gardener has six rows available in his vegetable garden to place tomatoes, eggplant, peppers, cucumbers, beans, and lettuce. Each vegetable will be allowed one and only one row. How many ways are there to position these vegetables in this garden?

4.47 There are eight members of a team. How many ways are there to select a team leader, assistant team leader, and team coordinator?

4.48 Four members of a group of 10 people are to be selected to a team. How many ways are there to select these four members?

4.49 A student has seven books that she would like to place in her backpack. However, there is room for only four books. Regardless of the arrangement, how many ways are there of placing four books into the backpack?

4.50 A daily lottery is conducted in which 2 winning numbers are selected out of 100 numbers. How many different combinations of winning numbers are possible?

4.51 A reading list for a course contains 20 articles. How many ways are there to choose 3 articles from this list?

4.5 Ethical Issues and Probability

Ethical issues can arise when any statements related to probability are presented to the public, particularly when these statements are part of an advertising campaign for a product or service. Unfortunately, many people are not comfortable with numerical concepts (see reference 5) and tend to misinterpret the meaning of the probability. In some instances, the misinterpretation is not intentional, but in other cases, advertisements may unethically try to mislead potential customers.

One example of a potentially unethical application of probability relates to advertisements for state lotteries. When purchasing a lottery ticket, the customer selects a set of numbers (such as 6) from a larger list of numbers (such as 54). Although virtually all participants know that they are unlikely to win the lottery, they also have very little idea of how unlikely it is for them to select all 6 winning numbers from the list of 54 numbers. They have even less of an idea of the probability of not selecting any winning numbers.

Given this background, you might consider a recent commercial for a state lottery that stated, “We won’t stop until we have made everyone a millionaire” to be deceptive and possibly unethical. Do you think the state has any intention of ever stopping the lottery, given the fact that the state relies on it to bring millions of dollars into its treasury? Is it possible that the lottery can make everyone a millionaire? Is it ethical to suggest that the purpose of the lottery is to make everyone a millionaire?

Another example of a potentially unethical application of probability relates to an investment newsletter promising a 90% probability of a 20% annual return on investment. To make the claim in the newsletter an ethical one, the investment service needs to (a) explain the basis on which this probability estimate rests, (b) provide the probability statement in another format, such as 9 chances in 10, and (c) explain what happens to the investment in the 10% of the cases in which a 20% return is not achieved (e.g., is the entire investment lost?).

These are serious ethical issues. If you were going to write an advertisement for the state lottery that ethically describes the probability of winning a certain prize, what would you say? If you were going to write an advertisement for the investment newsletter that ethically states the probability of a 20% return on an investment, what would you say?

USING STATISTICS

Possibilities at M&R Electronics World, Revisited

As the marketing manager for M&R Electronics World, you analyzed the survey results of an intent-to-purchase study. This study asked the heads of 1,000 households about their intentions to purchase a large-screen HDTV sometime during the next 12 months, and as a follow-up, M&R surveyed the same people 12 months later to see whether such a television was purchased. In addition, for households purchasing large-screen HDTVs, the survey asked whether the television they purchased had a faster refresh rate, whether they also purchased a streaming media box in the past 12 months, and whether they were satisfied with their purchase of the large-screen HDTV.

By analyzing the results of these surveys, you were able to uncover many pieces of valuable information that will help you plan a marketing strategy to enhance sales and better target those households likely to purchase multiple or more expensive products. Whereas only 30% of the households actually purchased a large-screen HDTV, if a household indicated that it planned to purchase a large-screen HDTV in the next 12 months, there was an 80% chance that the household actually made the purchase. Thus the marketing strategy



Shock/Fotolia

should target those households that have indicated an intention to purchase.

You determined that for households that purchased a television that had a faster refresh rate, there was a 47.5% chance that the household also purchased a streaming media box. You then compared this conditional probability to the marginal probability of purchasing a streaming media box, which was 36%. Thus, households that purchased televisions that had a faster refresh rate are more likely to purchase a streaming media box than are households that purchased large-screen HDTVs that have a standard refresh rate.

You were also able to apply Bayes' theorem to M&R Electronics World's market research reports. The reports investigate a potential new television model prior to its scheduled release. If a favorable report was received, then there was a 64% chance that the new television model would be successful. However, if an unfavorable report was received, there is only a 16% chance that the model would be successful. Therefore, the marketing strategy of M&R needs to pay close attention to whether a report's conclusion is favorable or unfavorable.

SUMMARY

This chapter began by developing the basic concepts of probability. You learned that probability is a numeric value from 0 to 1 that represents the chance, likelihood, or possibility that a particular event will occur. In addition to simple probability, you learned about conditional probabilities and independent events. Bayes' theorem was used to revise

previously calculated probabilities based on new information. Throughout the chapter, contingency tables and decision trees were used to display information. You also learned about several counting rules. In the next chapter, important discrete probability distributions such as the binomial, Poisson, and hypergeometric distributions are developed.

REFERENCES

1. Bellhouse, D. R. "The Reverend Thomas Bayes, FRS: A Biography to Celebrate the Tercentenary of His Birth." *Statistical Science*, 19 (2004), 3–43.
2. Lowd, D., and C. Meek. "Good Word Attacks on Statistical Spam Filters." Presented at the Second Conference on Email and Anti-Spam, 2005.
3. Microsoft Excel 2013. Redmond, WA: Microsoft Corp., 2012.
4. Minitab Release 16. State College, PA: Minitab, Inc, 2010.
5. Paulos, J. A. *Innumeracy*. New York: Hill and Wang, 1988.
6. Silberman, S. "The Quest for Meaning," *Wired* 8.02, February 2000.
7. Zeller, T. "The Fight Against V1@gra (and Other Spam)." *The New York Times*, May 21, 2006, pp. B1, B6.

KEY EQUATIONS

Probability of Occurrence

$$\text{Probability of occurrence} = \frac{X}{T} \quad (4.1)$$

Marginal Probability

$$P(A) = P(A \text{ and } B_1) + P(A \text{ and } B_2) + \dots + P(A \text{ and } B_k) \quad (4.2)$$

General Addition Rule

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \quad (4.3)$$

Conditional Probability

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \quad (4.4a)$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \quad (4.4b)$$

Independence

$$P(A|B) = P(A) \quad (4.5)$$

General Multiplication Rule

$$P(A \text{ and } B) = P(A|B)P(B) \quad (4.6)$$

Multiplication Rule for Independent Events

$$P(A \text{ and } B) = P(A)P(B) \quad (4.7)$$

Marginal Probability Using the General Multiplication Rule

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_k)P(B_k) \quad (4.8)$$

Bayes' Theorem

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_k)P(B_k)} \quad (4.9)$$

Counting Rule 1

$$k^n \quad (4.10)$$

Counting Rule 2

$$(k_1)(k_2) \dots (k_n) \quad (4.11)$$

Counting Rule 3

$$n! = (n)(n-1)\dots(1) \quad (4.12)$$

Counting Rule 4: Permutations

$${}_nP_x = \frac{n!}{(n-x)!} \quad (4.13)$$

Counting Rule 5: Combinations

$${}_nC_x = \frac{n!}{x!(n-x)!} \quad (4.14)$$

KEY TERMS

a priori probability 153

Bayes' theorem 169

certain event 152

collectively exhaustive 157

combination 176

complement 154

conditional probability 161

contingency table 155

decision tree 163

empirical probability 153

event 153

general addition rule 158

general multiplication rule 166

impossible event 152

independence 165

joint event 154

joint probability 156

marginal probability 157

multiplication rule for independent events 166

mutually exclusive 157

permutation 175

probability 152

sample space 154

simple event 153

simple probability 155

subjective probability 153

Venn diagram 155

CHECKING YOUR UNDERSTANDING

- 4.52** What are the differences between *a priori* probability, empirical probability, and subjective probability?
- 4.53** What is the difference between a simple event and a joint event?
- 4.54** How can you use the general addition rule to find the probability of occurrence of event *A* or *B*?
- 4.55** What is the difference between mutually exclusive events and collectively exhaustive events?
- 4.56** How does conditional probability relate to the concept of independence?
- 4.57** How does the multiplication rule differ for events that are and are not independent?
- 4.58** How can you use Bayes' theorem to revise probabilities in light of new information?
- 4.59** In Bayes' theorem, how does the prior probability differ from the revised probability?

CHAPTER REVIEW PROBLEMS

4.60 A survey by the Health Research Institute at PricewaterhouseCoopers LLP indicated that 80% of “young invincibles” (those aged 18 to 24) are likely to share health information through social media, as compared to 45% of “baby boomers” (those aged 45 to 64).

Source: Data extracted from “Social Media ‘Likes’ Healthcare: From Marketing to Social Business,” Health Research Institute, April 2012, p. 8.

Suppose that the survey was based on 500 respondents from each of the two groups.

- Construct a contingency table.
- Give an example of a simple event and a joint event.
- What is the probability that a randomly selected respondent is likely to share health information through social media?
- What is the probability that a randomly selected respondent is likely to share health information through social media *and* is in the 45-to-64-year-old group?
- Are the events “age group” and “likely to share health information through social media” independent? Explain.

4.61 SHL Americas provides a unique, global perspective of how talent is measured in its Global Assessment Trends Report. The report presents the results of an online survey conducted in late 2012 with HR professionals from companies headquartered throughout the world. The authors were interested in examining differences between respondents in *emerging economies* and those in *established economies* to provide relevant information for readers who may be creating assessment programs for organizations with global reach; one area of focus was on HR professionals’ response to two statements: “My organization views HR as a strategic function” and “My organization uses talent information to make business decisions.” The results are as follows:

ORGANIZATION VIEWS HR AS A STRATEGIC FUNCTION

ECONOMY	Yes	No	Total
Established	171	78	249
Emerging	222	121	343
Total	393	199	592

ORGANIZATION USES INFORMATION TALENT TO MAKE BUSINESS DECISIONS

ECONOMY	Yes	No	Total
Established	122	127	249
Emerging	130	213	343
Total	252	340	592

What is the probability that a randomly chosen HR professional

- is from an established economy?
- is from an established economy *or* agrees to the statement “My organization uses information talent to make business decisions?”
- does not agree with the statement “My organization views HR as a strategic function” *and* is from an emerging economy?
- does not agree with the statement “My organization views HR as a strategic function” *or* is from an emerging economy?
- Suppose the randomly chosen HR professional does not agree with the statement “My organization views HR as a strategic function.” What is the probability that the HR professional is from an emerging economy?
- Are “My organization views HR as a strategic function” and the type of economy independent?
- Is “My organization uses information talent to make business decisions” independent of the type of economy?

4.62 The 2012 Restaurant Industry Forecast takes a closer look at today’s consumers. Based on a 2011 National Restaurant Association survey, consumers are divided into three segments (optimistic, cautious, and hunkered-down) based on their financial situation, current spending behavior, and economic outlook. Suppose the results, based on a sample of 100 males and 100 females, were as follows:

GENDER

CONSUMER SEGMENT	Male	Female	Total
Optimistic	26	16	42
Cautious	41	43	84
Hunkered-down	33	41	74
Total	100	100	200

Source: Data extracted from “The 2012 Restaurant Industry Forecast,” National Restaurant Association, 2012, p. 12, restaurant.org/research/forecast.

If a consumer is selected at random, what is the probability that he or she

- is classified as cautious?
- is classified as optimistic or cautious?
- is a male *or* is classified as hunkered-down?
- is a male *and* is classified as hunkered-down?
- Given that the consumer selected is a female, what is the probability that she is classified as optimistic?

4.63 A 2011 joint study by MIT Sloan Management Review and IBM Institute for Business Value reports a growing divide between those companies that are transforming themselves to take advantage of business analytics and those that have yet to embrace it. A survey of business executives, managers, and analysts from organizations around the world indicated that 31% of organizations are “aspirational users” (basic users of analytics), 45% are “experienced users” (moderate users of analytics), and 24% are “transformed users” (strong and sophisticated users of analytics). Furthermore, 62% of the transformed-user organizations indicated an intense level of focus on using analytics to better understand and connect with customers, as did 49% of the experienced-user organizations and 34% of the aspirational-user organizations. (Data extracted from “Analytics: The Widening Divide, How Companies Are Achieving Competitive Advantage Through Analytics,” IBM Global Business Services, October, 2011.) If an organization is known to have an intense level of focus on using analytics to better understand and connect with customers, what is the probability that the organization is a transformed-user organization?

4.64 The CMO Council and SAS set out to better understand the key challenges, opportunities, and requirements that both chief marketing officers (CMOs) and chief information officers (CIOs) were facing in their journey to develop a more customer-centric enterprise. The following findings are from an online audit of 237 senior marketers and 210 senior IT executives. (Data extracted from “Big Data’s Biggest Role: Aligning the CMO & CIO,” March 2013, bit.ly/11z7uKW.)

BIG DATA IS CRITICAL TO EXECUTING A CUSTOMER-CENTRIC PROGRAM

EXECUTIVE GROUP	Yes	No	Total
Marketing	95	142	237
IT	107	103	210
Total	202	245	447

FUNCTIONAL SILOS BLOCK AGGREGATION OF CUSTOMER DATA THROUGHOUT THE ORGANIZATION

EXECUTIVE GROUP	Yes	No	Total
Marketing	122	115	237
IT	95	115	210
Total	217	230	447

- What is the probability that a randomly selected executive identifies Big Data as critical to executing a customer-centric program?

- Given that a randomly selected executive is a senior marketing executive, what is the probability that the executive identifies Big Data as critical to executing a customer-centric program?
- Given that a randomly selected executive is a senior IT executive, what is the probability that the executive identifies Big Data as critical to executing a customer-centric program?
- What is the probability that a randomly selected executive identifies that functional silos block aggregation of customer data throughout the organization?
- Given that a randomly selected executive is a senior marketing executive, what is the probability that the executive identifies that functional silos block aggregation of customer data throughout the organization?
- Given that a randomly selected executive is a senior IT executive, what is the probability that the executive identifies that functional silos block aggregation of customer data throughout the organization?
- Comment on the results in (a) through (f).

4.65 A 2013 Sage North America survey examined the “financial literacy” of small business owners. The study found that 23% of small business owners indicated concern about income tax compliance for their business; 41% of small business owners use accounting software, given that the small business owner indicated concern about income tax compliance for his or her business. Given that a small business owner did not indicate concern about income tax compliance for his or her business, 58% of small business owners use accounting software. (Data extracted from “Sage Financial Capability Survey: What Small Business Owners Don’t Understand Could Be Holding Them Back,” April 17, 2013, <http://bit.ly/Z3FAqx>.)

- Use Bayes’ theorem to find the probability that a small business owner uses accounting software, given that the small business owner indicated concern about income tax compliance for his or her business.
- Compare the result in (a) to the probability that a small business owner uses accounting software and comment on whether small business owners who are concerned about income tax compliance for their business are generally more likely to use accounting software than small business owners who are not concerned about income tax compliance for their business.

CASES FOR CHAPTER 4

Digital Case

Apply your knowledge about contingency tables and the proper application of simple and joint probabilities in this continuing Digital Case from Chapter 3.

Open **EndRunGuide.pdf**, the EndRun Financial Services “Guide to Investing,” and read the information about the Guaranteed Investment Package (GIP). Read the claims and examine the supporting data. Then answer the following questions:

1. How accurate is the claim of the probability of success for EndRun’s GIP? In what ways is the claim

misleading? How would you calculate and state the probability of having an annual rate of return not less than 15%?

2. Using the table found under the “Show Me the Winning Probabilities” subhead, compute the proper probabilities for the group of investors. What mistake was made in reporting the 7% probability claim?
3. Are there any probability calculations that would be appropriate for rating an investment service? Why or why not?

CardioGood Fitness

1. For each CardioGood Fitness treadmill product line (see the **CardioGoodFitness** file), construct two-way contingency tables of gender, education in years, relationship status, and self-rated fitness. (There will be a total of six tables for each treadmill product.)

2. For each table you construct, compute all conditional and marginal probabilities.
3. Write a report detailing your findings to be presented to the management of CardioGood Fitness.

The Choice Is Yours Follow-Up

1. Follow up the “Using Statistics: The Choice Is Yours, Revisited” on page 75 by constructing contingency tables of market cap and type, market cap and risk, market cap and rating, type and risk, type and rating, and risk and rating for the sample of 316 retirement funds stored in **Retirement Funds**.

2. For each table you construct, compute all conditional and marginal probabilities.
3. Write a report summarizing your conclusions.

Clear Mountain State Student Surveys

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receive responses from 62 undergraduates (stored in **UndergradSurvey**).

1. For these data, construct contingency tables of gender and major, gender and graduate school intention, gender and employment status, gender and computer preference, class and graduate school intention, class and employment status, major and graduate school intention, major and employment status, and major and computer preference.
 - a. For each of these contingency tables, compute all the conditional and marginal probabilities.
 - b. Write a report summarizing your conclusions.

2. The CMSU Dean of Students has learned about the undergraduate survey and has decided to undertake a similar survey for graduate students at Clear Mountain State. She creates and distributes a survey of 14 questions and receives responses from 44 graduate students (stored in **GradSurvey**). Construct contingency tables of gender and graduate major, gender and undergraduate major, gender and employment status, gender and computer preference, graduate major and undergraduate major, graduate major and employment status, and graduate major and computer preference.

- a. For each of these contingency tables, compute all the conditional and marginal probabilities.
- b. Write a report summarizing your conclusions.

CHAPTER 4 EXCEL GUIDE

EG4.1 BASIC PROBABILITY CONCEPTS

Simple and Joint Probability and the General Addition Rule

Key Technique Use Excel arithmetic formulas.

Example Compute simple and joint probabilities for the Table 4.1 purchase behavior data on page 154.

PHStat2 Use Simple & Joint Probabilities.

For the example, select **PHStat → Probability & Prob. Distributions → Simple & Joint Probabilities**. In the new template, similar to the worksheet shown below, fill in the **Sample Space** area with the data.

In-Depth Excel Use the **COMPUTE worksheet** of the **Probabilities workbook** as a template.

The worksheet (shown below) already contains the Table 4.1 purchase behavior data. For other problems, change the sample space table entries in the cell ranges **C3:D4** and **A5:D6**.

A	B	C	D	E	
1	Probabilities				
2					
3	Sample Space	Actually Purchased			
4		Yes	No	Totals	
5	Planned to Purchase	Yes	200	50	250
6		No	100	650	750
7	Totals	300	700	1000	
8					
9	Simple Probabilities				
10	P(Yes)	0.25	=E5/E7		
11	P(No)	0.75	=E6/E7		
12	P(Yes)	0.30	=C7/E7		
13	P(No)	0.70	=D7/E7		
14					
15	Joint Probabilities				
16	P(Yes and Yes)	0.20	=C5/E7		
17	P(Yes and No)	0.05	=D5/E7		
18	P(No and Yes)	0.10	=C6/E7		
19	P(No and No)	0.65	=D6/E7		
20					
21	Addition Rule				
22	P(Yes or Yes)	0.35	=H16 + H18 - H22		
23	P(Yes or No)	0.90	=H16 + H19 - H23		
24	P(No or Yes)	0.95	=H17 + H18 - H24		
25	P(No or No)	0.80	=H17 + H19 - H25		

Read the **SHORT TAKES** for Chapter 4 for an explanation of the formulas found in the **COMPUTE worksheet** (shown in the **COMPUTE_FORMULAS worksheet**).

EG4.2 CONDITIONAL PROBABILITY

There is no Excel material for this section.

EG4.3 BAYES' THEOREM

Key Technique Use Excel arithmetic formulas.

Example Apply Bayes' theorem to the television marketing example in Section 4.3.

In-Depth Excel Use the **COMPUTE worksheet** of the **Bayes workbook** as a template.

The worksheet (shown below) already contains the probabilities for the Section 4.3 example. For other problems, change those probabilities in the cell range **B5:C6**.

A	B	C	D	E	
1 Bayes' Theorem Computations					
2					
3					
4	Event	Prior	Conditional	Joint	Revised
5	S	0.4	0.8	0.32	0.64
6	S'	0.6	0.3	0.18	0.36
7		Total: 0.5		Joint	Revised
			=B5 * C5	=D5/\$D\$7	
			=B6 * C6	=D6/\$D\$7	
			=D5 + D6		

Open to the **COMPUTE_FORMULAS worksheet** to examine the arithmetic formulas that compute the probabilities, which are also shown as an inset to the worksheet.

EG4.4 COUNTING RULES

Counting Rule 1

In-Depth Excel Use the **POWER(*k, n*)** worksheet function in a cell formula to compute the number of outcomes given *k* events and *n* trials. For example, the formula **=POWER(6, 2)** computes the answer for Example 4.11 on page 174.

Counting Rule 2

In-Depth Excel Use a formula that takes the product of successive **POWER(*k, n*)** functions to solve problems related to counting rule 2. For example, the formula **=POWER(26, 3) * POWER(10, 3)** computes the answer for the state motor vehicle department example on page 174.

Counting Rule 3

In-Depth Excel Use the **FACT(*n*)** worksheet function in a cell formula to compute how many ways *n* items can be arranged. For example, the formula **=FACT(6)** computes 6!

Counting Rule 4

In-Depth Excel Use the **PERMUT(*n, x*)** worksheet function in a cell formula to compute the number of ways of arranging *x* objects selected from *n* objects in order. For example, the formula **=PERMUT(6, 4)** computes the answer for Example 4.14 on page 176.

Counting Rule 5

In-Depth Excel Use the **COMBIN(*n, x*)** worksheet function in a cell formula to compute the number of ways of arranging *x* objects selected from *n* objects, irrespective of order. For example, the formula **=COMBIN(6, 4)** computes the answer for Example 4.15 on page 176.

CHAPTER 4 MINITAB GUIDE

MG4.1 BASIC PROBABILITY CONCEPTS

There is no Minitab material for this section.

MG4.2 CONDITIONAL PROBABILITY

There is no Minitab material for this section.

MG4.3 BAYES' THEOREM

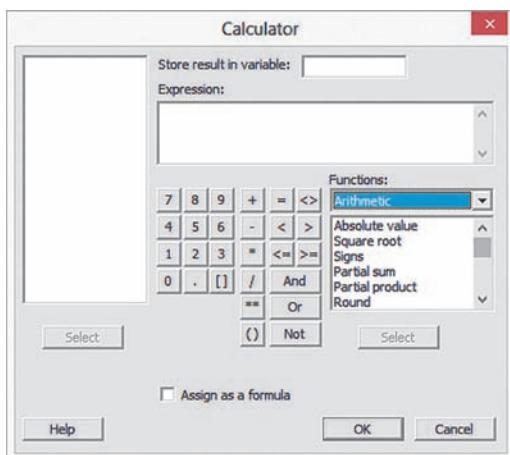
There is no Minitab material for this section.

MG4.4 COUNTING RULES

Use **Calculator** to apply the counting rules. Select **Calc → Calculator**. In the Calculator dialog box (shown below):

1. Enter the column name of an empty column in the **Store result in variable** box and then press **Tab**.
2. Build the appropriate expression (as discussed later in this section) in the **Expression** box. To apply counting rules 3 through 5, select **Arithmetic** from the **Functions** drop-down list to facilitate the function selection.
3. Click **OK**.

If you have previously used the Calculator during your Minitab session, you may have to clear the contents of the Expression box by selecting the contents and pressing **Del** before you begin step 2.



Counting Rule 1

Enter an expression that uses the exponential operator ******. For example, the expression **6 ** 2** computes the answer for Example 4.11 on page 174.

Counting Rule 2

Enter an expression that uses the exponential operator ******. For example, the expression **26 ** 3 * 10 ** 3** computes the answer for the state motor vehicle department example on page 174.

Counting Rule 3

Enter an expression that uses the **FACTORIAL(n)** function to compute how many ways n items can be arranged. For example, the expression **FACTORIAL(6)** computes $6!$

Counting Rule 4

Enter an expression that uses the **PERMUTATIONS(n, x)** function to compute the number of ways of arranging x objects selected from n objects in order. For example, the expression **PERMUTATIONS(6, 4)** computes the answer for Example 4.14 on page 176.

Counting Rule 5

Enter an expression that uses the **COMBINATIONS(n, x)** function to compute the number of ways of arranging x objects selected from n objects, irrespective of order. For example, the expression **COMBINATIONS(6, 4)** computes the answer for Example 4.15 on page 176.

CHAPTER 5

Discrete Probability Distributions

CONTENTS

- 5.1 The Probability Distribution for a Discrete Variable
- 5.2 Covariance of a Probability Distribution and Its Application in Finance
- 5.3 Binomial Distribution
- 5.4 Poisson Distribution
- 5.5 Hypergeometric Distribution
- 5.6 Using the Poisson Distribution to Approximate the Binomial Distribution (*online*)

USING STATISTICS: Events of Interest at Ricknel Home Centers, Revisited

CHAPTER 5 EXCEL GUIDE

CHAPTER 5 MINITAB GUIDE

OBJECTIVES

- To learn the properties of a probability distribution
- To compute the expected value and variance of a probability distribution
- To calculate the covariance and understand its use in finance
- To compute probabilities from the binomial, Poisson, and hypergeometric distributions
- To use the binomial, Poisson, and hypergeometric distributions to solve business problems

USING STATISTICS

Events of Interest at Ricknel Home Centers

Like most other large businesses, Ricknel Home Centers, LLC, a regional home improvement chain, uses an accounting information system (AIS) to manage its accounting and financial data. The Ricknel AIS collects, organizes, stores, analyzes, and distributes financial information to decision makers both inside and outside the firm.

One important function of the Ricknel AIS is to continuously audit accounting information, looking for errors or incomplete or improbable information. For example, when customers submit orders online, the Ricknel AIS reviews the orders for possible mistakes. Any questionable invoices are tagged and included in a daily *exceptions report*. Recent data collected by the company show that the likelihood is 0.10 that an order form will be tagged.

As a member of the AIS team, you have been asked by Ricknel management to determine the likelihood of finding a certain number of tagged forms in a sample of a specific size. For example, what would be the likelihood that none of the order forms are tagged in a sample of four forms? That one of the order forms is tagged?

How could you determine the solution to this type of probability problem?



Sebastian Kaulitzki/Shutterstock

This chapter introduces you to the concept and characteristics of probability distributions. You will learn how the binomial, Poisson, and hypergeometric distributions can be applied to help solve business problems. In the Rickel Home Centers scenario, you could use a *probability distribution* as a mathematical model, or small-scale representation, that approximates the process. By using such an approximation, you could make inferences about the actual order process including the likelihood of finding a certain number of tagged forms in a sample.

5.1 The Probability Distribution for a Discrete Variable

Recall from Section 1.1 that *numerical* variables are variables that have values that represent quantities, such as the one-year return percentage for a retirement fund or the number of social media sites to which you belong. Some numerical variables are *discrete*, having numerical values that arise from a counting process, while others are *continuous*, having numerical values that arise from a measuring process (e.g., the one-year return of growth and value funds that were the subject of the Using Statistics scenario in Chapters 2 and 3). This chapter deals with probability distributions that represent a discrete numerical variable, such as the number of social media sites to which you belong.

PROBABILITY DISTRIBUTION FOR A DISCRETE VARIABLE

A **probability distribution for a discrete variable** is a mutually exclusive list of all the possible numerical outcomes along with the probability of occurrence of each outcome.

For example, Table 5.1 gives the distribution of the number of interruptions per day in a large computer network. The list in Table 5.1 is collectively exhaustive because all possible outcomes are included. Thus, the probabilities sum to 1. Figure 5.1 is a graphical representation of Table 5.1.

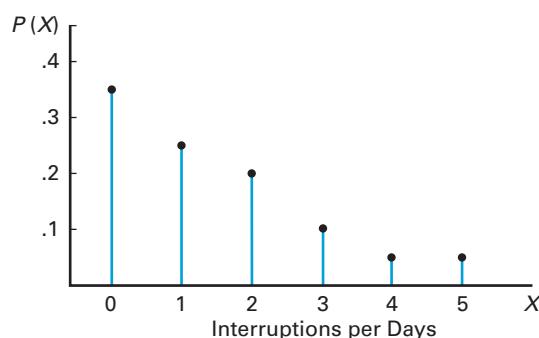
TABLE 5.1

Probability Distribution of the Number of Interruptions per Day

Interruptions per Day	Probability
0	0.35
1	0.25
2	0.20
3	0.10
4	0.05
5	0.05

FIGURE 5.1

Probability distribution of the number of interruptions per day



Student Tip

Remember, *expected value* is just the *mean*.

Expected Value of a Discrete Variable

The **expected value** of a random variable is the mean, μ , of its probability distribution. To calculate the expected value, you multiply each possible outcome, x_i , by its corresponding probability, $P(X = x_i)$, and then sum these products.

EXPECTED VALUE, μ , OF A DISCRETE VARIABLE

$$\mu = E(X) = \sum_{i=1}^N x_i P(X = x_i) \quad (5.1)$$

where

x_i = the i th value of the discrete variable X

$P(X = x_i)$ = probability of occurrence of the i th value of X

For the probability distribution of the number of interruptions per day in a large computer network (Table 5.1), the expected value is computed as follows, using Equation (5.1), and is also shown in Table 5.2:

$$\begin{aligned}\mu &= E(X) = \sum_{i=1}^N x_i P(X = x_i) \\ &= (0)(0.35) + (1)(0.25) + (2)(0.20) + (3)(0.10) + (4)(0.05) + (5)(0.05) \\ &= 0 + 0.25 + 0.40 + 0.30 + 0.20 + 0.25 \\ &= 1.40\end{aligned}$$

TABLE 5.2

Computing the Expected Value of the Number of Interruptions per Day

Interruptions per Day (x_i)	$P(X = x_i)$	$x_i P(X = x_i)$
0	0.35	(0)(0.35) = 0.00
1	0.25	(1)(0.25) = 0.25
2	0.20	(2)(0.20) = 0.40
3	0.10	(3)(0.10) = 0.30
4	0.05	(4)(0.05) = 0.20
5	0.05	(5)(0.05) = 0.25
	1.00	$\mu = E(X) = 1.40$

The expected value is 1.40. The expected value of 1.40 interruptions per day is not a possible result because the actual number of interruptions on a given day must be an integer value. The expected value represents the *mean* number of interruptions on a given day.

Variance and Standard Deviation of a Discrete Variable

You compute the variance of a probability distribution by multiplying each possible squared difference $[x_i - E(X)]^2$ by its corresponding probability, $P(X = x_i)$, and then summing the resulting products. Equation (5.2) defines the **variance of a discrete variable**, and Equation (5.3) defines the **standard deviation of a discrete variable**.

VARIANCE OF A DISCRETE VARIABLE

$$\sigma^2 = \sum_{i=1}^N [x_i - E(X)]^2 P(X = x_i) \quad (5.2)$$

where

x_i = the i th value of the discrete variable X

$P(X = x_i)$ = probability of occurrence of the i th value of X

Use the Section EG5.1 instructions to compute the variance and standard deviation of a discrete variable.

STANDARD DEVIATION OF A DISCRETE VARIABLE

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^N [x_i - E(X)]^2 P(X = x_i)} \quad (5.3)$$

The variance and the standard deviation of the number of interruptions per day are computed as follows and in Table 5.3, using Equations (5.2) and (5.3):

$$\begin{aligned}\sigma^2 &= \sum_{i=1}^N [x_i - E(X)]^2 P(X = x_i) \\ &= (0 - 1.4)^2(0.35) + (1 - 1.4)^2(0.25) + (2 - 1.4)^2(0.20) + (3 - 1.4)^2(0.10) \\ &\quad + (4 - 1.4)^2(0.05) + (5 - 1.4)^2(0.05) \\ &= 0.686 + 0.040 + 0.072 + 0.256 + 0.338 + 0.648 \\ &= 2.04\end{aligned}$$

and

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.04} = 1.4283$$

TABLE 5.3

Computing the Variance and Standard Deviation of the Number of Interruptions per Day

Interruptions per Day (x_i)	$P(X = x_i)$	$x_i P(X = x_i)$	$[x_i - E(X)]^2$	$[x_i - E(X)]^2 P(X = x_i)$
0	0.35	0.00	$(0 - 1.4)^2 = 1.96$	$(1.96)(0.35) = 0.686$
1	0.25	0.25	$(1 - 1.4)^2 = 0.16$	$(0.16)(0.25) = 0.040$
2	0.20	0.40	$(2 - 1.4)^2 = 0.36$	$(0.36)(0.20) = 0.072$
3	0.10	0.30	$(3 - 1.4)^2 = 2.56$	$(2.56)(0.10) = 0.256$
4	0.05	0.20	$(4 - 1.4)^2 = 6.76$	$(6.76)(0.05) = 0.338$
5	0.05	0.25	$(5 - 1.4)^2 = 12.96$	$(12.96)(0.05) = 0.648$
	1.00	$\mu = E(X) = 1.40$		$\sigma^2 = 2.04$
				$\sigma = \sqrt{\sigma^2} = 1.4283$

Thus, the mean number of interruptions per day is 1.4, the variance is 2.04, and the standard deviation is approximately 1.43 interruptions per day.

Problems for Section 5.1

LEARNING THE BASICS

5.1 Given the following probability distributions:

- a. Compute the expected value for each distribution.
- b. Compute the standard deviation for each distribution.
- c. Compare the results of distributions A and B.

Distribution A		Distribution B	
x_i	$P(X = x_i)$	x_i	$P(X = x_i)$
0	0.50	0	0.05
1	0.20	1	0.10
2	0.15	2	0.15
3	0.10	3	0.20
4	0.05	4	0.50

APPLYING THE CONCEPTS

 SELF Test

- 5.2** The following table contains the probability distribution for the number of traffic accidents daily in a small town:

Number of Accidents Daily (X)	$P(X = x_i)$
0	0.10
1	0.20
2	0.45
3	0.15
4	0.05
5	0.05

- a. Compute the mean number of accidents per day.
 b. Compute the standard deviation.
- 5.3** Recently, a regional automobile dealership sent out fliers to perspective customers indicating that they had already won one of three different prizes: an automobile valued at \$25,000, a \$100 gas card, or a \$5 Walmart shopping card. To claim his or her prize, a prospective customer needed to present the flier at the dealership's showroom. The fine print on the back of the flier listed the probabilities of winning. The chance of winning the car was 1 out of 31,478, the chance of winning the gas card was 1 out of 31,478, and the chance of winning the shopping card was 31,476 out of 31,478.
- a. How many fliers do you think the automobile dealership sent out?
 - b. Using your answer to (a) and the probabilities listed on the flier, what is the expected value of the prize won by a prospective customer receiving a flier?
 - c. Using your answer to (a) and the probabilities listed on the flier, what is the standard deviation of the value of the prize won by a prospective customer receiving a flier?
 - d. Do you think this is an effective promotion? Why or why not?

- 5.4** In the carnival game Under-or-Over-Seven, a pair of fair dice is rolled once, and the resulting sum determines whether the player wins or loses his or her bet. For example, the player can bet \$1 that the sum will be under 7—that is, 2, 3, 4, 5, or 6. For this bet, the player wins \$1 if the result is under 7 and loses \$1 if the outcome equals or is greater than 7. Similarly, the player can bet \$1 that the sum will be over 7—that is, 8, 9, 10, 11, or 12. Here, the player wins \$1 if the result is over 7 but loses \$1 if the result is 7 or under. A third method of play is to bet \$1 on the outcome 7. For this bet, the player wins \$4 if the result of the roll is 7 and loses \$1 otherwise.

- a. Construct the probability distribution representing the different outcomes that are possible for a \$1 bet on under 7.

- b. Construct the probability distribution representing the different outcomes that are possible for a \$1 bet on over 7.
- c. Construct the probability distribution representing the different outcomes that are possible for a \$1 bet on 7.
- d. Show that the expected long-run profit (or loss) to the player is the same, no matter which method of play is used.

- 5.5** The number of arrivals per minute at a bank located in the central business district of a large city was recorded over a period of 200 minutes, with the following results:

Arrivals	Frequency
0	14
1	31
2	47
3	41
4	29
5	21
6	10
7	5
8	2

- a. Compute the expected number of arrivals per minute.
- b. Compute the standard deviation.

- 5.6** The manager of the commercial mortgage department of a large bank has collected data during the past two years concerning the number of commercial mortgages approved per week. The results from these two years (104 weeks) are as follows:

Number of Commercial Mortgages Approved	Frequency
0	13
1	25
2	32
3	17
4	9
5	6
6	1
7	1

- a. Compute the expected number of mortgages approved per week.
- b. Compute the standard deviation.

5.2 Covariance of a Probability Distribution and Its Application in Finance

Section 5.1 defined the expected value, variance, and standard deviation for a single discrete variable. In this section, the covariance between two variables is introduced and applied to portfolio management, a topic of great interest to financial analysts.

Covariance

The **covariance of a probability distribution** (σ_{XY}) measures the strength of the relationship between two variables, X and Y . A positive covariance indicates a positive relationship. A negative covariance indicates a negative relationship. If two variables are independent, their covariance will be zero. Equation (5.4) defines the covariance of discrete random variables X and Y .

COVARIANCE

$$\sigma_{XY} = \sum_{i=1}^N [x_i - E(X)][y_i - E(Y)]P(x_i, y_i) \quad (5.4)$$

where

X = discrete variable X

x_i = i th value of X

Y = discrete variable Y

y_i = i th value of Y

$P(x_i, y_i)$ = probability of occurrence of the i th value of X and the i th value of Y

$i = 1, 2, \dots, N$ for X and Y

To illustrate the covariance, suppose that you are deciding between two different investments for the coming year. The first investment is a mutual fund that consists of the stocks that comprise the Dow Jones Industrial Average. The second investment is a mutual fund that is expected to perform best when economic conditions are weak. Table 5.4 summarizes your estimate of the returns (per \$1,000 investment) under three economic conditions, each with a given probability of occurrence.

TABLE 5.4

Estimated Returns
for Each Investment
Under Three
Economic Conditions

$P(x_i, y_i)$	Economic Condition	Investment Return	
		Dow Jones Fund	Weak-Economy Fund
0.2	Recession	-\$300	+\$200
0.5	Stable economy	+100	+50
0.3	Expanding economy	+250	-100

The expected value and standard deviation for each investment and the covariance of the two investments are computed as follows:

Student Tip

The covariance discussed in this section measures the strength of the linear relationship between the *probability distributions* of two variables, while the *sample covariance* discussed in Chapter 3 measures the strength of the linear relationship between two numerical variables.

Let X = the return of the Dow Jones fund and Y = the return of the weak-economy fund

$$E(X) = \mu_X = (-300)(0.2) + (100)(0.5) + (250)(0.3) = \$65$$

$$E(Y) = \mu_Y = (+200)(0.2) + (50)(0.5) + (-100)(0.3) = \$35$$

$$\begin{aligned} Var(X) &= \sigma_X^2 = (-300 - 65)^2(0.2) + (100 - 65)^2(0.5) + (250 - 65)^2(0.3) \\ &= 37,525 \end{aligned}$$

$$\sigma_X = \$193.71$$

$$\begin{aligned} Var(Y) &= \sigma_Y^2 = (200 - 35)^2(0.2) + (50 - 35)^2(0.5) + (-100 - 35)^2(0.3) \\ &= 11,025 \end{aligned}$$

$$\sigma_Y = \$105.00$$

$$\begin{aligned} \sigma_{XY} &= (-300 - 65)(200 - 35)(0.2) + (100 - 65)(50 - 35)(0.5) \\ &\quad + (250 - 65)(-100 - 35)(0.3) \\ &= -12,045 + 262.5 - 7,492.5 \\ &= -19,275 \end{aligned}$$

Thus, the Dow Jones fund has a higher expected value (i.e., larger expected return) than the weak-economy fund but also has a higher standard deviation (i.e., more risk). The covariance of $-19,275$ between the two investment returns indicates a negative relationship in which the return of two investments are varying in the *opposite* direction. Therefore, when the return on one investment is high, typically, the return on the other investment is low.

Expected Value, Variance, and Standard Deviation of the Sum of Two Variables

Equations (5.1) through (5.3) define the expected value, variance, and standard deviation of a probability distribution, and Equation (5.4) defines the covariance between two variables, X and Y . The **expected value of the sum of two variables** is equal to the sum of the expected values. The **variance of the sum of two variables** is equal to the sum of the variances plus twice the covariance. The **standard deviation of the sum of two variables** is the square root of the variance of the sum of two variables.

EXPECTED VALUE OF THE SUM OF TWO VARIABLES

$$E(X + Y) = E(X) + E(Y) \quad (5.5)$$

VARIANCE OF THE SUM OF TWO VARIABLES

$$Var(X + Y) = \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} \quad (5.6)$$

STANDARD DEVIATION OF THE SUM OF TWO VARIABLES

$$\sigma_{X+Y} = \sqrt{\sigma_{X+Y}^2} \quad (5.7)$$

To illustrate the expected value, variance, and standard deviation of the sum of two variables, consider the two investments previously discussed. If X = return of the Dow Jones fund and Y = return of the weak-economy fund, using Equations (5.5), (5.6), and (5.7),

$$\begin{aligned} E(X + Y) &= E(X) + E(Y) = 65 + 35 = \$100 \\ \sigma_{X+Y}^2 &= \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} \\ &= 37,525 + 11,025 + (2)(-19,275) \\ &= 10,000 \\ \sigma_{X+Y} &= \$100 \end{aligned}$$

The expected value of the sum of the return of the Dow Jones fund and the return of the weak-economy fund is \$100, with a standard deviation of \$100. The standard deviation of the sum of the two investments is less than the standard deviation of either single investment because there is a large negative covariance between the investments.

Portfolio Expected Return and Portfolio Risk

The covariance and the expected value and standard deviation of the sum of two random variables can be applied to analyzing **portfolios**, or groupings of assets made for investment purposes. Investors combine assets into portfolios to reduce their risk (see references 1 and 2). Often, the objective is to try to maximize the return while making the risk as small as possible. For such portfolios, rather than study the sum of two random variables, the investor weights each investment by the proportion of assets assigned to that investment. Equations (5.8) and (5.9) define the **portfolio expected return** and **portfolio risk**.

PORTFOLIO EXPECTED RETURN

The portfolio expected return for a two-asset investment is equal to the weight assigned to asset X multiplied by the expected return of asset X plus the weight assigned to asset Y multiplied by the expected return of asset Y .

$$E(P) = wE(X) + (1 - w)E(Y) \quad (5.8)$$

where

$E(P)$ = portfolio expected return

w = portion of the portfolio value assigned to asset X

$(1 - w)$ = portion of the portfolio value assigned to asset Y

$E(X)$ = expected return of asset X

$E(Y)$ = expected return of asset Y

PORTFOLIO RISK

The portfolio risk for a two-asset investment is equal to the square root of the sum of these three products: w^2 multiplied by the variance of X , $(1 - w)^2$ multiplied by the variance of Y , and 2 multiplied by w multiplied by $(1 - w)$ multiplied by the covariance.

$$\sigma_p = \sqrt{w^2\sigma_X^2 + (1 - w)^2\sigma_Y^2 + 2w(1 - w)\sigma_{XY}} \quad (5.9)$$

In the previous section, you evaluated the expected return and risk of two different investments, a Dow Jones fund and a weak-economy fund. You also computed the covariance of the two investments. Now, suppose that you want to form a portfolio of these two investments that consists of an equal investment in each of these two funds. To compute the portfolio expected return and the portfolio risk, using Equations (5.8) and (5.9), with $w = 0.50$, $E(X) = \$65$, $E(Y) = \$35$, $\sigma_X^2 = 37,525$, $\sigma_Y^2 = 11,025$, and $\sigma_{XY} = -19,275$,

$$E(P) = (0.5)(65) + (1 - 0.5)(35) = \$50$$

$$\begin{aligned}\sigma_p &= \sqrt{(0.5)^2(37,525) + (1 - 0.5)^2(11,025) + 2(0.5)(1 - 0.5)(-19,275)} \\ &= \sqrt{2,500} = \$50\end{aligned}$$

Thus, the portfolio has an expected return of \$50 for each \$1,000 invested (a return of 5%) and a portfolio risk of \$50. The portfolio risk here is smaller than the standard deviation of either investment because there is a large negative covariance between the two investments. The fact that each investment performs best under different circumstances reduces the overall risk of the portfolio.

Collapses in the financial marketplace that have occurred in the recent past have caused some investors to consider the effect of outcomes that have only a small chance of occurring but that could produce extremely negative results. (Some, including the author of reference 6, have labeled these outcomes “black swans.”) Example 5.1 considers such an outcome by examining the expected return, the standard deviation of the return, and the covariance of two investment strategies—one that invests in a fund that does well when there is an extreme recession and the other that invests in a fund that does well under positive economic conditions.

EXAMPLE 5.1

Computing the Expected Return, the Standard Deviation of the Return, and the Covariance of Two Investment Strategies

You plan to invest \$1,000 in one of two funds. Table 5.5 shows the annual return (per \$1,000) of each of these investments under different economic conditions, along with the probability that each of these economic conditions will occur.

TABLE 5.5

Estimated Returns of Two Funds

Probability	Economic Condition	Black Swan Fund	Good Times Fund
0.01	Extreme recession	400	-200
0.09	Recession	-30	-100
0.15	Stagnation	30	50
0.35	Slow growth	50	90
0.30	Moderate growth	100	250
0.10	High growth	100	225

For the Black Swan fund and the Good Times fund, compute the expected return and standard deviation of the return for each fund, and the covariance between the two funds. Would you invest in the Black Swan fund or the Good Times fund? Explain.

SOLUTION Let X = Black Swan fund and Y = Good Times fund.

$$\begin{aligned} E(X) &= \mu_X = (400)(0.01) + (-30)(0.09) + (30)(0.15) + (50)(0.35) \\ &\quad + (100)(0.30) + (100)(0.10) = \$63.30 \end{aligned}$$

$$\begin{aligned} E(Y) &= \mu_Y = (-200)(0.01) + (-100)(0.09) + (50)(0.15) + (90)(0.35) \\ &\quad + (250)(0.30) + (225)(0.10) = \$125.50 \end{aligned}$$

$$\begin{aligned} Var(X) &= \sigma_X^2 = (400 - 63.30)^2(0.01) + (-30 - 63.30)^2(0.09) + (30 - 63.30)^2(0.15) \\ &\quad + (50 - 63.30)^2(0.35) + (100 - 63.30)^2(0.3) + (100 - 63.30)^2(0.1) = 2,684.11 \\ \sigma_X &= \$51.81 \end{aligned}$$

$$\begin{aligned} Var(Y) &= \sigma_Y^2 = (-200 - 125.50)^2(0.01) + (-100 - 125.50)^2(0.09) \\ &\quad + (50 - 125.50)^2(0.15) + (90 - 125.50)^2(0.35) + (250 - 125.50)^2(0.3) \\ &\quad + (225 - 125.50)^2(0.1) = 12,572.25 \end{aligned}$$

$$\sigma_Y = \$112.13$$

$$\begin{aligned} \sigma_{XY} &= (400 - 63.30)(-200 - 125.50)(0.01) + (-30 - 63.30)(-100 - 125.50)(0.09) \\ &\quad + (30 - 63.30)(50 - 125.50)(0.15) + (50 - 63.30)(90 - 125.50)(0.35) \\ &\quad + (100 - 63.30)(250 - 125.50)(0.3) + (100 - 63.30)(225 - 125.50)(0.1) \\ \sigma_{xy} &= \$3,075.85 \end{aligned}$$

Thus, the Good Times fund not only has a much higher expected value (i.e., larger expected return) than the Black Swan fund (\$125.50 as compared to \$63.30 per \$1,000) but also has a much higher standard deviation (\$112.13 vs. \$51.81). Deciding which fund to invest in is a matter of how much risk you are willing to tolerate. Although the Good Times fund has a much higher expected return, many people would be reluctant to invest in a fund where there is a chance of a substantial loss.

The covariance of \$3,075.85 between the two investments indicates a positive relationship in which the two investments are varying in the *same* direction. Therefore, when the return on one investment is high, typically, the return on the other is also high. However, from Table 5.5, you can see that the magnitude of the return varies, depending on the economic condition that actually occurs. Therefore, you might decide to include both funds in your portfolio. The percentage allocated to each fund would be based on your tolerance of risk balanced by your desire for maximum return (see Problem 5.15).

Problems for Section 5.2

LEARNING THE BASICS

5.7 Given the following probability distributions for variables X and Y :

$P(x,y)$	X	Y
0.4	100	200
0.6	200	100

Compute

- a. $E(X)$ and $E(Y)$.
- b. σ_X and σ_Y .
- c. σ_{XY} .
- d. $E(X + Y)$.

5.8 Given the following probability distributions for variables X and Y :

$P(x,y)$	X	Y
0.2	-100	50
0.4	50	30
0.3	200	20
0.1	300	20

Compute

- a. $E(X)$ and $E(Y)$.
- b. σ_X and σ_Y .
- c. σ_{XY} .
- d. $E(X + Y)$.

5.9 Two investments, X and Y , have the following characteristics:

$$E(X) = \$50, E(Y) = \$100, \sigma_X^2 = 9,000, \\ \sigma_Y^2 = 15,000, \text{ and } \sigma_{XY} = 7,500.$$

If the weight of portfolio assets assigned to investment X is 0.4, compute the

- a. portfolio expected return.
- b. portfolio risk.

APPLYING THE CONCEPTS

5.10 The process of being served at a bank consists of two independent parts—the time waiting in line and the time it takes to be served by the teller. Suppose that the time waiting in line has an expected value of 4 minutes, with a standard deviation of 1.2 minutes, and the time it takes to be served by the teller has an expected value of 5.5 minutes, with a standard deviation of 1.5 minutes. Compute the

- a. expected value of the total time it takes to be served at the bank.
- b. standard deviation of the total time it takes to be served at the bank.

5.11 In the portfolio example in this section (see page 192), half the portfolio assets are invested in the Dow Jones fund and half in a weak-economy fund. Recalculate the portfolio expected return and the portfolio risk if

- a. 30% of the portfolio assets are invested in the Dow Jones fund and 70% in a weak-economy fund.

- b. 70% of the portfolio assets are invested in the Dow Jones fund and 30% in a weak-economy fund.

- c. Which of the three investment strategies (30%, 50%, or 70% in the Dow Jones fund) would you recommend? Why?

 **5.12** You are trying to develop a strategy for investing in two different stocks. The anticipated annual return for a \$1,000 investment in each stock under four different economic conditions has the following probability distribution:

Probability	Economic Condition	Returns	
		Stock X	Stock Y
0.1	Recession	-100	50
0.3	Slow growth	0	150
0.3	Moderate growth	80	-20
0.3	Fast growth	150	-100

Compute the

- a. expected return for stock X and for stock Y .
- b. standard deviation for stock X and for stock Y .
- c. covariance of stock X and stock Y .
- d. Would you invest in stock X or stock Y ? Explain.

5.13 Suppose that in Problem 5.12 you wanted to create a portfolio that consists of stock X and stock Y . Compute the portfolio expected return and portfolio risk for each of the following percentages invested in stock X :

- a. 30%
- b. 50%
- c. 70%
- d. On the basis of the results of (a) through (c), which portfolio would you recommend? Explain.

5.14 You are trying to develop a strategy for investing in two different stocks. The anticipated annual return for a \$1,000 investment in each stock under four different economic conditions has the following probability distribution:

Probability	Economic Condition	Returns	
		Stock X	Stock Y
0.1	Recession	-50	-100
0.3	Slow growth	20	50
0.4	Moderate growth	100	130
0.2	Fast growth	150	200

Compute the

- a. expected return for stock X and for stock Y .
- b. standard deviation for stock X and for stock Y .
- c. covariance of stock X and stock Y .
- d. Would you invest in stock X or stock Y ? Explain.

5.15 Suppose that in Example 5.1 on page 193, you wanted to create a portfolio that consists of the Black Swan fund and the Good Times fund. Compute the portfolio expected return and portfolio risk for each of the following percentages invested in the Black Swan fund:

- 30%
- 50%
- 70%
- On the basis of the results of (a) through (c), which portfolio would you recommend? Explain.

5.16 You plan to invest \$1,000 in a corporate bond fund or in a common stock fund. The following table presents the annual return (per \$1,000) of each of these investments under various economic conditions and the probability that each of those economic conditions will occur. Compute the

Probability	Economic Condition	Corporate Bond Fund	Common Stock Fund
0.01	Extreme recession	-200	-999
0.09	Recession	-70	-300
0.15	Stagnation	30	-100
0.35	Slow growth	80	100
0.30	Moderate growth	100	150
0.10	High growth	120	350

- expected return for the corporate bond fund and for the common stock fund.
- standard deviation for the corporate bond fund and for the common stock fund.
- covariance of the corporate bond fund and the common stock fund.
- Would you invest in the corporate bond fund or the common stock fund? Explain.
- If you chose to invest in the common stock fund in (d), what do you think about the possibility of losing \$999 of every \$1,000 invested if there is an extreme recession?

5.17 Suppose that in Problem 5.16 you wanted to create a portfolio that consists of the corporate bond fund and the common stock fund. Compute the portfolio expected return and portfolio risk for each of the following situations:

- \$300 in the corporate bond fund and \$700 in the common stock fund.
- \$500 in each fund.
- \$700 in the corporate bond fund and \$300 in the common stock fund.
- On the basis of the results of (a) through (c), which portfolio would you recommend? Explain.

5.3 Binomial Distribution

This is the first of three sections that considers mathematical models. A **mathematical model** is a mathematical expression that represents a variable of interest. When a mathematical model exists, you can compute the exact probability of occurrence of any particular value of the variable. For discrete random variables, the mathematical model is a **probability distribution function**.

The **binomial distribution** is an important mathematical model used in many business situations. You use the binomial distribution when the discrete variable is the number of events of interest in a sample of n observations. The binomial distribution has four important properties.

Student Tip

Do not confuse this use of the Greek letter pi, π , to represent the probability of an event of interest with the mathematical constant that is the ratio of the circumference to a diameter of a circle—approximately 3.14159—which is also known by the same Greek letter.

PROPERTIES OF THE BINOMIAL DISTRIBUTION

- The sample consists of a fixed number of observations, n .
- Each observation is classified into one of two mutually exclusive and collectively exhaustive categories.
- The probability of an observation being classified as the event of interest, π , is constant from observation to observation. Thus, the probability of an observation being classified as not being the event of interest, $1 - \pi$, is constant over all observations.
- The value of any observation is independent of the value of any other observation.

Returning to the Ricknel Home Improvement scenario presented on page 185 concerning the accounting information system, suppose the event of interest is defined as a tagged order form. You want to determine the number of tagged order forms in a given sample of orders.

What results can occur? If the sample contains four orders, there could be none, one, two, three, or four tagged order forms. No other value can occur because the number of tagged

order forms cannot be more than the sample size, n , and cannot be less than zero. Therefore, the range of the binomial random variable is from 0 to n .

Suppose that you observe the following result in a sample of four orders:

First Order	Second Order	Third Order	Fourth Order
Tagged	Tagged	Not tagged	Tagged

What is the probability of having three tagged order forms in a sample of four orders in this particular sequence? Because the historical probability of a tagged order is 0.10, the probability that each order occurs in the sequence is

First Order	Second Order	Third Order	Fourth Order
$\pi = 0.10$	$\pi = 0.10$	$1 - \pi = 0.90$	$\pi = 0.10$

Each outcome is independent of the others because the order forms were selected from an extremely large or practically infinite population and each order form could only be selected once. Therefore, the probability of having this particular sequence is

$$\begin{aligned}\pi\pi(1 - \pi)\pi &= \pi^3(1 - \pi)^1 \\ &= (0.10)^3(0.90)^1 \\ &= (0.10)(0.10)(0.10)(0.90) \\ &= 0.0009\end{aligned}$$

This result indicates only the probability of three tagged order forms (events of interest) from a sample of four order forms in a *specific sequence*. To find the number of ways of selecting x objects from n objects, *irrespective of sequence*, you use the **rule of combinations**¹ given in Equation (5.10).

¹Refer to Section 4.4 for further discussion of counting rules.

²On many scientific calculators, there is a button labeled $_nC_r$ that allows you to compute the number of combinations. On these calculators, the symbol r is used instead of x .

COMBINATIONS

The number of combinations of selecting x objects² out of n objects is given by

$${}_nC_x = \frac{n!}{x!(n-x)!} \quad (5.10)$$

where

$n! = (n)(n - 1) \cdots (1)$ is called n factorial. By definition, $0! = 1$.

With $n = 4$ and $x = 3$, there are

$${}_nC_x = \frac{n!}{x!(n-x)!} = \frac{4!}{3!(4-3)!} = \frac{4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)(1)} = 4$$

such sequences. The four possible sequences are

Sequence 1 = (*tagged, tagged, tagged, not tagged*), with probability
 $\pi\pi\pi(1 - \pi) = \pi^3(1 - \pi)^1 = 0.0009$

Sequence 2 = (*tagged, tagged, not tagged, tagged*), with probability
 $\pi\pi(1 - \pi)\pi = \pi^3(1 - \pi)^1 = 0.0009$

Sequence 3 = (*tagged, not tagged, tagged, tagged*), with probability
 $\pi(1 - \pi)\pi\pi = \pi^3(1 - \pi)^1 = 0.0009$

Sequence 4 = (*not tagged, tagged, tagged, tagged*), with probability
 $(1 - \pi)\pi\pi\pi = \pi^3(1 - \pi)^1 = 0.0009$

Therefore, the probability of three tagged order forms is equal to

$$\begin{aligned} & (\text{number of possible sequences}) \times (\text{probability of a particular sequence}) \\ & = (4) \times (0.0009) = 0.0036 \end{aligned}$$

You can make a similar, intuitive derivation for the other possible values of the random variable—zero, one, two, and four tagged order forms. However, as n , the sample size, gets large, the computations involved in using this intuitive approach become time-consuming. Equation (5.11) is the mathematical model that provides a general formula for computing any probability from the binomial distribution with the number of events of interest, x , given n and π .

BINOMIAL DISTRIBUTION

$$P(X = x | n, \pi) = \frac{n!}{x!(n-x)!} \pi^x (1 - \pi)^{n-x} \quad (5.11)$$

where

$P(X = x | n, \pi)$ = probability that $X = x$ events of interest, given n and π

n = number of observations

π = probability of an event of interest

$1 - \pi$ = probability of not having an event of interest

x = number of events of interest in the sample ($X = 0, 1, 2, \dots, n$)

$\frac{n!}{x!(n-x)!}$ = number of combinations of x events of interest out of n observations

Equation (5.11) restates what was intuitively derived previously. The binomial variable X can have any integer value x from 0 through n . In Equation (5.11), the product

$$\pi^x (1 - \pi)^{n-x}$$

represents the probability of exactly x events of interest from n observations in a *particular sequence*.

The term

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

is the number of *combinations* of the x events of interest from the n observations possible. Hence, given the number of observations, n , and the probability of an event of interest, π , the probability of x events of interest is

$$\begin{aligned} P(X = x | n, \pi) &= (\text{number of combinations}) \times (\text{probability of a particular combination}) \\ &= \frac{n!}{x!(n-x)!} \pi^x (1 - \pi)^{n-x} \end{aligned}$$

Example 5.2 illustrates the use of Equation (5.11). Examples 5.3 and 5.4 show the computations for other values of X .

EXAMPLE 5.2

Determining
 $P(X = 3)$, Given
 $n = 4$ and $\pi = 0.1$

If the likelihood of a tagged order form is 0.1, what is the probability that there are three tagged order forms in the sample of four?

SOLUTION Using Equation (5.11), the probability of three tagged orders from a sample of four is

$$\begin{aligned} P(X = 3 | n = 4, \pi = 0.1) &= \frac{4!}{3!(4-3)!}(0.1)^3(1-0.1)^{4-3} \\ &= \frac{4!}{3!(1)!}(0.1)^3(0.9)^1 \\ &= 4(0.1)(0.1)(0.1)(0.9) = 0.0036 \end{aligned}$$

EXAMPLE 5.3

Determining
 $P(X \geq 3)$, Given
 $n = 4$ and $\pi = 0.1$

If the likelihood of a tagged order form is 0.1, what is the probability that there are three or more (i.e., at least three) tagged order forms in the sample of four?

SOLUTION In Example 5.2, you found that the probability of *exactly* three tagged order forms from a sample of four is 0.0036. To compute the probability of *at least* three tagged order forms, you need to add the probability of three tagged order forms to the probability of four tagged order forms. The probability of four tagged order forms is

$$\begin{aligned} P(X = 4 | n = 4, \pi = 0.1) &= \frac{4!}{4!(4-4)!}(0.1)^4(1-0.1)^{4-4} \\ &= \frac{4!}{4!(0)!}(0.1)^4(0.9)^0 \\ &= 1(0.1)(0.1)(0.1)(0.1)(1) = 0.0001 \end{aligned}$$

Student Tip

Another way of saying "three or more" is "at least three."

Thus, the probability of at least three tagged order forms is

$$\begin{aligned} P(X \geq 3) &= P(X = 3) + P(X = 4) \\ &= 0.0036 + 0.0001 \\ &= 0.0037 \end{aligned}$$

There is a 0.37% chance that there will be at least three tagged order forms in a sample of four.

EXAMPLE 5.4

Determining
 $P(X < 3)$, Given
 $n = 4$ and $\pi = 0.1$

If the likelihood of a tagged order form is 0.1, what is the probability that there are less than three tagged order forms in the sample of four?

SOLUTION The probability that there are less than three tagged order forms is

$$P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2)$$

Using Equation (5.11) on page 197, these probabilities are

$$P(X = 0 | n = 4, \pi = 0.1) = \frac{4!}{0!(4-0)!}(0.1)^0(1-0.1)^{4-0} = 0.6561$$

$$P(X = 1 | n = 4, \pi = 0.1) = \frac{4!}{1!(4-1)!} (0.1)^1 (1 - 0.1)^{4-1} = 0.2916$$

$$P(X = 2 | n = 4, \pi = 0.1) = \frac{4!}{2!(4-2)!} (0.1)^2 (1 - 0.1)^{4-2} = 0.0486$$

Therefore, $P(X < 3) = 0.6561 + 0.2916 + 0.0486 = 0.9963$. $P(X < 3)$ could also be calculated from its complement, $P(X \geq 3)$, as follows:

$$\begin{aligned} P(X < 3) &= 1 - P(X \geq 3) \\ &= 1 - 0.0037 = 0.9963 \end{aligned}$$

Computing binomial probabilities become tedious as n gets large. Figure 5.2 shows how Excel and Minitab can compute binomial probabilities for you. You can also look up binomial probabilities in a table of probabilities.

FIGURE 5.2

Excel and Minitab results for computing binomial probabilities with $n = 4$ and $\pi = 0.1$

A		B	Cumulative Distribution Function
1 Binomial Probabilities			Binomial with n = 4 and p = 0.1
2			x P(X <= x)
3 Data			0 0.6561
4 Sample size		4	1 0.9477
5 Probability of an event of interest		0.1	2 0.9963
6			3 0.9999
7 Statistics			4 1.0000
8 Mean		0.4 =B4 * B5	
9 Variance		0.36 =B8 * (1 - B5)	
10 Standard deviation		0.6 =SQRT(B9)	
11			
12 Binomial Probabilities Table			
13 X		P(X)	
14 0		0.6561 =BINOM.DIST(A14, \$B\$4, \$B\$5, FALSE)	
15 1		0.2916 =BINOM.DIST(A15, \$B\$4, \$B\$5, FALSE)	
16 2		0.0486 =BINOM.DIST(A16, \$B\$4, \$B\$5, FALSE)	
17 3		0.0036 =BINOM.DIST(A17, \$B\$4, \$B\$5, FALSE)	
18 4		0.0001 =BINOM.DIST(A18, \$B\$4, \$B\$5, FALSE)	

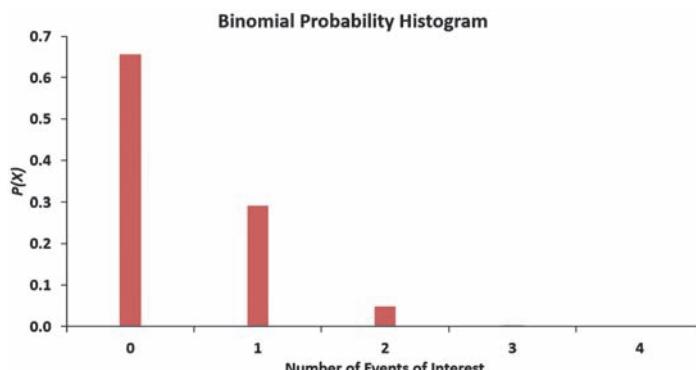
LEARN MORE

The **Binomial Table** online topic contains both a binomial probabilities table and a cumulative binomial probabilities table and explains how to use these tables to compute binomial and cumulative binomial probabilities.

The shape of a binomial probability distribution depends on the values of n and π . Whenever $\pi = 0.5$, the binomial distribution is symmetrical, regardless of how large or small the value of n . When $\pi \neq 0.5$, the distribution is skewed. The closer π is to 0.5 and the larger the number of observations, n , the less skewed the distribution becomes. For example, the distribution of the number of tagged order forms is highly right skewed because $\pi = 0.1$ and $n = 4$ (see Figure 5.3).

FIGURE 5.3

Histogram of the binomial probability with $n = 4$ and $\pi = 0.1$



Observe from Figure 5.3 that unlike the histogram for continuous variables in Section 2.4, the bars for the values are very thin, and there is a large gap between each pair of values. That is because the histogram represents a discrete variable. (Theoretically, the bars should have no width. They should be vertical lines.)

The mean (or expected value) of the binomial distribution is equal to the product of n and π . Instead of using Equation (5.1) on page 187 to compute the mean of the probability distribution, you can use Equation (5.12) to compute the mean for variables that follow the binomial distribution.

MEAN OF THE BINOMIAL DISTRIBUTION

The mean, μ , of the binomial distribution is equal to the sample size, n , multiplied by the probability of an event of interest, π .

$$\mu = E(X) = n\pi \quad (5.12)$$

On the average, over the long run, you theoretically expect $\mu = E(X) = n\pi] = (4)(0.1) = 0.4$ tagged order form in a sample of four orders.

The standard deviation of the binomial distribution can be calculated using Equation (5.13).

STANDARD DEVIATION OF THE BINOMIAL DISTRIBUTION

$$\sigma = \sqrt{\sigma^2} = \sqrt{Var(X)} = \sqrt{n\pi(1 - \pi)} \quad (5.13)$$

The standard deviation of the number of tagged order forms is

$$\sigma = \sqrt{4(0.1)(0.9)} = 0.60$$

You get the same result if you use Equation (5.3) on page 188.

Example 5.5 applies the binomial distribution to service at a fast-food restaurant.

EXAMPLE 5.5

Computing Binomial Probabilities for Service at a Fast-Food Restaurant

Accuracy in taking orders at a drive-through window is important for fast-food chains. Periodically, *QSR Magazine* publishes the results of a survey that measures accuracy, defined as the percentage of orders that are filled correctly. In a recent year, the percentage of orders filled correctly at Wendy's was approximately 88.9% (bit.ly/QEIeOW). Suppose that you go to the drive-through window at Wendy's and place an order. Two friends of yours independently place orders at the drive-through window at the same Wendy's. What are the probabilities that all three, that none of the three, and that at least two of the three orders will be filled correctly? What are the mean and standard deviation of the binomial distribution for the number of orders filled correctly?

SOLUTION Because there are three orders and the probability of a correct order is 0.889, $n = 3$, and $\pi = 0.889$, using Equation (5.11) on page 197,

$$\begin{aligned} P(X = 3 | n = 3, \pi = 0.889) &= \frac{3!}{3!(3-3)!}(0.889)^3(1 - 0.889)^{3-3} \\ &= \frac{3!}{3!(3-3)!}(0.889)^3(0.111)^0 \\ &= 1(0.889)(0.889)(0.889)(1) = 0.7026 \end{aligned}$$

$$P(X = 0 | n = 3, \pi = 0.889) = \frac{3!}{0!(3-0)!} (0.889)^0 (1 - 0.889)^{3-0}$$

$$= \frac{3!}{0!(3-0)!} (0.889)^0 (0.111)^3$$

$$= 1(1)(0.111)(0.111)(0.111) = 0.0014$$

$$P(X = 2 | n = 3, \pi = 0.889) = \frac{3!}{2!(3-2)!} (0.889)^2 (1 - 0.889)^{3-2}$$

$$= \frac{3!}{2!(3-2)!} (0.889)^2 (0.111)^1$$

$$= 3(0.889)(0.889)(0.111) = 0.2632$$

$$\begin{aligned} P(X \geq 2) &= P(X = 2) + P(X = 3) \\ &= 0.2632 + 0.7026 \\ &= 0.9658 \end{aligned}$$

Using Equations (5.12) and (5.13),

$$\begin{aligned} \mu &= E(X) = n\pi = 3(0.889) = 2.667 \\ \sigma &= \sqrt{\sigma^2} = \sqrt{Var(X)} = \sqrt{n\pi(1 - \pi)} \\ &= \sqrt{3(0.889)(0.111)} \\ &= \sqrt{0.2960} = 0.5441 \end{aligned}$$

The mean number of orders filled correctly in a sample of three orders is 2.667, and the standard deviation is 0.5441. The probability that all three orders are filled correctly is 0.7026, or 70.26%. The probability that none of the orders are filled correctly is 0.0014, or 0.14%. The probability that at least two orders are filled correctly is 0.9658, or 96.58%.

Problems for Section 5.3

LEARNING THE BASICS

5.18 Determine the following:

- a. For $n = 4$ and $\pi = 0.12$, what is $P(X = 0)$?
- b. For $n = 10$ and $\pi = 0.40$, what is $P(X = 9)$?
- c. For $n = 10$ and $\pi = 0.50$, what is $P(X = 8)$?
- d. For $n = 6$ and $\pi = 0.83$, what is $P(X = 5)$?

5.19 If $n = 5$ and $\pi = 0.40$, what is the probability that

- | | |
|---------------|------------|
| a. $X = 4$ | c. $X < 2$ |
| b. $X \leq 3$ | d. $X > 1$ |

5.20 Determine the mean and standard deviation of the variable X in each of the following binomial distributions:

- a. $n = 4$ and $\pi = 0.10$
- b. $n = 4$ and $\pi = 0.40$
- c. $n = 5$ and $\pi = 0.80$
- d. $n = 3$ and $\pi = 0.50$

APPLYING THE CONCEPTS

5.21 The increase or decrease in the price of a stock between the beginning and the end of a trading day is assumed to be an equally likely random event. What is the probability that a stock will show an increase in its closing price on five consecutive days?

5.22 A recent YouGov (UK) survey reported that 27% of under-25-year-olds in the United Kingdom own tablets. (Data extracted from “Tablets Spur News Consumption,” bit.ly/12XpmR0). Using the binomial distribution, what is the probability that in the next six under-25-year-olds surveyed,

- a. four will own a tablet?
- b. all six will own a tablet?
- c. at least four will own a tablet?
- d. What are the mean and standard deviation of the number of under-25-year-olds who will own a tablet in a survey of six?
- e. What assumptions do you need to make in (a) through (c)?

5.23 A student is taking a multiple-choice exam in which each question has four choices. Assume that the student has no knowledge of the correct answers to any of the questions. She has decided on a strategy in which she will place four balls (marked *A*, *B*, *C*, and *D*) into a box. She randomly selects one ball for each question and replaces the ball in the box. The marking on the ball will determine her answer to the question. There are five multiple-choice questions on the exam. What is the probability that she will get

- five questions correct?
- at least four questions correct?
- no questions correct?
- no more than two questions correct?

5.24 A manufacturing company regularly conducts quality control checks at specified periods on the products it manufactures. Historically, the failure rate for LED light bulbs that the company manufactures is 5%. Suppose a random sample of 10 LED light bulbs is selected. What is the probability that

- none of the LED light bulbs are defective?
- exactly one of the LED light bulbs is defective?
- two or fewer of the LED light bulbs are defective?
- three or more of the LED light bulbs are defective?

5.25 When a customer places an order with Rudy's On-Line Office Supplies, a computerized accounting information system (AIS) automatically checks to see if the customer has exceeded his or her credit limit. Past records indicate that the probability of customers exceeding their credit limit is 0.05. Suppose that, on a given day, 20 customers place orders. Assume that the number of customers that the AIS detects as having exceeded their credit limit is distributed as a binomial random variable.

- What are the mean and standard deviation of the number of customers exceeding their credit limits?
- What is the probability that zero customers will exceed their credit limits?
- What is the probability that one customer will exceed his or her credit limit?
- What is the probability that two or more customers will exceed their credit limits?



5.26 In Example 5.5 on page 200, you and two friends decided to go to Wendy's. Now, suppose that instead you go to Burger King, which recently filled approximately 83% of orders correctly. What is the probability that

- all three orders will be filled correctly?
- none of the three will be filled correctly?
- at least two of the three will be filled correctly?
- What are the mean and standard deviation of the binomial distribution used in (a) through (c)? Interpret these values.

5.27 In Example 5.5 on page 200, you and two friends decided to go to Wendy's. Now, suppose that instead you go to McDonald's, which recently filled approximately 90.9% of the orders correctly. What is the probability that

- all three orders will be filled correctly?
- none of the three will be filled correctly?
- at least two of the three will be filled correctly?
- What are the mean and standard deviation of the binomial distribution used in (a) through (c)? Interpret these values.
- Compare the result of (a) through (d) with those of Burger King in Problem 5.26 and Wendy's in Example 5.5 on page 200.

5.4 Poisson Distribution

Many studies are based on counts of the occurrences of a particular event in a given interval of time or space (often referred to as an *area of opportunity*). In such an **area of opportunity** there can be more than one occurrence of an event. The Poisson distribution can be used to compute probabilities in such situations. Examples of variables that follow the Poisson distribution are the surface defects on a new refrigerator, the number of network failures in a day, the number of people arriving at a bank, and the number of fleas on the body of a dog. You can use the **Poisson distribution** to calculate probabilities in situations such as these if the following properties hold:

- You are interested in counting the number of times a particular event occurs in a given area of opportunity. The area of opportunity is defined by time, length, surface area, and so forth.
- The probability that an event occurs in a given area of opportunity is the same for all the areas of opportunity.
- The number of events that occur in one area of opportunity is independent of the number of events that occur in any other area of opportunity.
- The probability that two or more events will occur in an area of opportunity approaches zero as the area of opportunity becomes smaller.

Consider the number of customers arriving during the lunch hour at a bank located in the central business district in a large city. You are interested in the number of customers who arrive each minute. Does this situation match the four properties of the Poisson distribution given earlier?

First, the *event* of interest is a customer arriving, and the *given area of opportunity* is defined as a one-minute interval. Will zero customers arrive, one customer arrive, two customers arrive, and so on? Second, it is reasonable to assume that the probability that a customer arrives during a particular one-minute interval is the same as the probability for all the other one-minute intervals. Third, the arrival of one customer in any one-minute interval has no effect on (i.e., is independent of) the arrival of any other customer in any other one-minute interval. Finally, the probability that two or more customers will arrive in a given time period approaches zero as the time interval becomes small. For example, the probability is virtually zero that two customers will arrive in a time interval of 0.01 second. Thus, you can use the Poisson distribution to determine probabilities involving the number of customers arriving at the bank in a one-minute time interval during the lunch hour.

The Poisson distribution has one characteristic, called λ (the Greek lowercase letter *lambda*), which is the mean or expected number of events per unit. The variance of a Poisson distribution is also equal to λ , and the standard deviation is equal to $\sqrt{\lambda}$. The number of events, X , of the Poisson random variable ranges from 0 to infinity (∞).

Equation (5.14) is the mathematical expression for the Poisson distribution for computing the probability of $X = x$ events, given that λ events are expected.

POISSON DISTRIBUTION

$$P(X = x | \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \quad (5.14)$$

where

$P(X = x | \lambda)$ = probability that $X = x$ events in an area of opportunity given λ

λ = expected number of events

e = mathematical constant approximated by 2.71828

x = number of events ($x = 0, 1, 2, \dots$)

To illustrate an application of the Poisson distribution, suppose that the mean number of customers who arrive per minute at the bank during the noon-to-1 P.M. hour is equal to 3.0. What is the probability that in a given minute, exactly two customers will arrive? And what is the probability that more than two customers will arrive in a given minute?

Using Equation (5.14) and $\lambda = 3$, the probability that in a given minute exactly two customers will arrive is

$$P(X = 2 | \lambda = 3) = \frac{e^{-3.0}(3.0)^2}{2!} = \frac{9}{(2.71828)^3(2)} = 0.2240$$

To determine the probability that in any given minute more than two customers will arrive,

$$P(X > 2) = P(X = 3) + P(X = 4) + \dots$$

Because in a probability distribution, all the probabilities must sum to 1, the terms on the right side of the equation $P(X > 2)$ also represent the complement of the probability that X is less than or equal to 2 [i.e., $1 - P(X \leq 2)$]. Thus,

$$P(X > 2) = 1 - P(X \leq 2) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$$

Now, using Equation (5.14),

$$\begin{aligned} P(X > 2) &= 1 - \left[\frac{e^{-3.0}(3.0)^0}{0!} + \frac{e^{-3.0}(3.0)^1}{1!} + \frac{e^{-3.0}(3.0)^2}{2!} \right] \\ &= 1 - [0.0498 + 0.1494 + 0.2240] \\ &= 1 - 0.4232 = 0.5768 \end{aligned}$$

Thus, there is a 57.68% chance that more than two customers will arrive in the same minute.

Computing Poisson probabilities can be tedious. Figure 5.4 shows how Excel and Minitab can compute Poisson probabilities for you. You can also look up Poisson probabilities in a table of probabilities.

FIGURE 5.4

Excel and Minitab results for computing Poisson probabilities with $\lambda = 3$

LEARN MORE
The [Poisson Table online topic](#) contains a table of Poisson probabilities and explains how to use the table to compute Poisson probabilities.

A	B	C	D	E
1	Poisson Probabilities			
2				
3	Data			
4	Mean/Expected number of events of interest:	3		
5				
6	Poisson Probabilities Table			
7	X	P(X)		
8	0	0.0498	=POISSON.DIST(A8, \$E\$4, FALSE)	
9	1	0.1494	=POISSON.DIST(A9, \$E\$4, FALSE)	
10	2	0.2240	=POISSON.DIST(A10, \$E\$4, FALSE)	
11	3	0.2240	=POISSON.DIST(A11, \$E\$4, FALSE)	
12	4	0.1680	=POISSON.DIST(A12, \$E\$4, FALSE)	
13	5	0.1008	=POISSON.DIST(A13, \$E\$4, FALSE)	
14	6	0.0504	=POISSON.DIST(A14, \$E\$4, FALSE)	
15	7	0.0216	=POISSON.DIST(A15, \$E\$4, FALSE)	
16	8	0.0081	=POISSON.DIST(A16, \$E\$4, FALSE)	
17	9	0.0027	=POISSON.DIST(A17, \$E\$4, FALSE)	
18	10	0.0008	=POISSON.DIST(A18, \$E\$4, FALSE)	
19	11	0.0002	=POISSON.DIST(A19, \$E\$4, FALSE)	
20	12	0.0001	=POISSON.DIST(A20, \$E\$4, FALSE)	
21	13	0.0000	=POISSON.DIST(A21, \$E\$4, FALSE)	
22	14	0.0000	=POISSON.DIST(A22, \$E\$4, FALSE)	
23	15	0.0000	=POISSON.DIST(A23, \$E\$4, FALSE)	

Probability Density Function	
Poisson with mean = 3	
x	P(X=x)
0	0.049787
1	0.149361
2	0.224042
3	0.224042
4	0.168031
5	0.100819
6	0.050409
7	0.021604
8	0.008102
9	0.002701
10	0.000810
11	0.000221
12	0.000055
13	0.000013
14	0.000003
15	0.000001

EXAMPLE 5.6

Computing Poisson Probabilities

The number of work-related injuries per month in a manufacturing plant is known to follow a Poisson distribution, with a mean of 2.5 work-related injuries a month. What is the probability that in a given month, no work-related injuries occur? That at least one work-related injury occurs?

SOLUTION Using Equation (5.14) on page 203 with $\lambda = 2.5$ (or Excel, Minitab, or a Poisson table lookup), the probability that in a given month no work-related injuries occur is

$$P(X = 0 | \lambda = 2.5) = \frac{e^{-2.5}(2.5)^0}{0!} = \frac{1}{(2.71828)^{2.5}(1)} = 0.0821$$

The probability that there will be no work-related injuries in a given month is 0.0821, or 8.21%. Thus,

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - 0.0821 \\ &= 0.9179 \end{aligned}$$

The probability that there will be at least one work-related injury is 0.9179, or 91.79%.

Problems for Section 5.4

LEARNING THE BASICS

5.28 Assume a Poisson distribution.

- a. If $\lambda = 2.5$, find $P(X = 2)$.
- b. If $\lambda = 8.0$, find $P(X = 8)$.
- c. If $\lambda = 0.5$, find $P(X = 1)$.
- d. If $\lambda = 3.7$, find $P(X = 0)$.

5.29 Assume a Poisson distribution.

- a. If $\lambda = 2.0$, find $P(X \geq 2)$.
- b. If $\lambda = 8.0$, find $P(X \geq 3)$.
- c. If $\lambda = 0.5$, find $P(X \leq 1)$.
- d. If $\lambda = 4.0$, find $P(X \geq 1)$.
- e. If $\lambda = 5.0$, find $P(X \leq 3)$.

5.30 Assume a Poisson distribution with $\lambda = 5.0$. What is the probability that

- | | |
|------------|---------------|
| a. $X = 1$ | c. $X > 1$ |
| b. $X < 1$ | d. $X \leq 1$ |

APPLYING THE CONCEPTS

5.31 Assume that the number of network errors experienced in a day on a local area network (LAN) is distributed as a Poisson variable. The mean number of network errors experienced in a day is 2.4. What is the probability that in any given day

- a. zero network errors will occur?
- b. exactly one network error will occur?
- c. two or more network errors will occur?
- d. fewer than three network errors will occur?



5.32 The quality control manager of Marilyn's Cookies is inspecting a batch of chocolate-chip cookies that has just been baked. If the production process is in control, the mean number of chocolate-chip parts per cookie is 6.0. What is the probability that in any particular cookie being inspected

- a. fewer than five chocolate-chip parts will be found?
- b. exactly five chocolate-chip parts will be found?
- c. five or more chocolate-chip parts will be found?
- d. either four or five chocolate-chip parts will be found?

5.33 Refer to Problem 5.32. How many cookies in a batch of 100 should the manager expect to discard if company policy requires that all chocolate-chip cookies sold have at least four chocolate-chip parts?

5.34 The U.S. Department of Transportation maintains statistics for mishandled bags per 1,000 airline passengers. In February 2013, Delta mishandled 2.05 bags per 1,000 passengers. What is the probability that in the next 1,000 passengers, Delta will have

- a. no mishandled bags?
- b. at least one mishandled bag?
- c. at least two mishandled bags?

5.35 The U.S. Department of Transportation maintains statistics for involuntary denial of boarding. In February 2013, the American Airlines rate of involuntarily denying boarding was 0.74 per 10,000 passengers. What is the probability that in the next 10,000 passengers, there will be

- a. no one involuntarily denied boarding?
- b. at least one person involuntarily denied boarding?
- c. at least two persons involuntarily denied boarding?

5.36 The Consumer Financial Protection Bureau's consumer response team hears directly from consumers about the challenges they face in the marketplace, brings their concerns to the attention of financial institutions, and assists in addressing their complaints. The consumer response team accepts complaints related to mortgages, bank accounts and services, private student loans, other consumer loans, and credit reporting. An analysis of complaints over time indicates that the mean number of credit reporting complaints registered by consumers is 2.15 per day. (Source: *Consumer Response: A Snapshot of Complaints Received, 1.usa.gov/WZ9N8Q*.) Assume that the number of credit reporting complaints registered by consumers is distributed as a Poisson random variable. What is the probability that on a given day

- a. no credit reporting complaints will be registered by consumers?
- b. exactly one credit reporting complaint will be registered by consumers?
- c. more than one credit reporting complaint will be registered by consumers?
- d. fewer than two credit reporting complaints will be registered by consumers?

5.37 J.D. Power and Associates calculates and publishes various statistics concerning car quality. The dependability score measures problems experienced during the past 12 months by original owners of three-year-old vehicles (those that were introduced for the 2010 model year). For these models of cars, Ford had 1.27 problems per car and Toyota had 1.12 problems per car. (Data extracted from "2013 U.S. Vehicle Dependability Study," J.D. Power and Associates, February 13, 2013, bit.ly/101aR9I.) Let X be equal to the number of problems with a three-year-old Ford.

- a. What assumptions must be made in order for X to be distributed as a Poisson random variable? Are these assumptions reasonable?

Making the assumptions as in (a), if you purchased a Ford in the 2010 model year, what is the probability that in the past 12 months, the car had

- b. zero problems?
- c. two or fewer problems?

d. Give an operational definition for *problem*. Why is the operational definition important in interpreting the initial quality score?

5.38 Refer to Problem 5.37. If you purchased a Toyota in the 2010 model year, what is the probability that in the past 12 months the car had

- a. zero problems?
- b. two or fewer problems?
- c. Compare your answers in (a) and (b) to those for the Ford in Problem 5.37 (b) and (c).

5.39 Refer to Problem 5.37. Another press release reported in 2012 that for 2009 model cars, Ford had 1.24 problems per car and Toyota had 1.04 problems per car. (Data extracted from "2013 U.S. Vehicle Dependability Study," J.D. Power and Associates, February 13, 2013, bit.ly/101aR9I.) If you purchased a 2009 Ford, what is the probability that in the past 12 months the car had

- a. zero problems?
- b. two or fewer problems?
- c. Compare your answers in (a) and (b) to those for the 2010 model year Ford in Problem 5.37 (b) and (c).

- 5.40** Refer to Problem 5.39. If you purchased a 2009 Toyota, what is the probability that in the past 12 months, the car had
- zero problems?
 - two or fewer problems?
 - Compare your answers in (a) and (b) to those for the 2010 model year Toyota in Problem 5.38 (a) and (b).

5.41 A toll-free phone number is available from 9 A.M. to 9 P.M. for your customers to register complaints about a product purchased from your company. Past history indicates that an average of 0.8 calls is received per minute.

- What properties must be true about the situation described here in order to use the Poisson distribution to calculate probabilities concerning the number of phone calls received in a one-minute period?

Assuming that this situation matches the properties discussed in (a), what is the probability that during a one-minute period

- zero phone calls will be received?
- three or more phone calls will be received?
- What is the maximum number of phone calls that will be received in a one-minute period 99.99% of the time?

5.5 Hypergeometric Distribution

Both the binomial distribution and the **hypergeometric distribution** use the number of events of interest in a sample containing n observations. One of the differences in these two probability distributions is in the way the samples are selected. For the binomial distribution, the sample data are selected *with replacement* from a *finite* population or *without replacement* from an *infinite* population. Thus, the probability of an event of interest, π , is constant over all observations, and the result of any particular observation is independent of the result of any other observation. For the hypergeometric distribution, the sample data are selected *without replacement* from a *finite* population. Thus, the result of one observation is dependent on the results of the previous observations.

Consider a population of size N . Let E represent the total number of events of interest in the population. The hypergeometric distribution is then used to find the probability of x events of interest in a sample of size n , selected without replacement. Equation (5.15) represents the mathematical expression of the hypergeometric distribution for finding x events of interest, given a knowledge of n , N , and E .

HYPERGEOMETRIC DISTRIBUTION

$$P(X = x | n, N, E) = \frac{\binom{E}{x} \binom{N - E}{n - x}}{\binom{N}{n}} \quad (5.15)$$

where

$P(X = x | n, N, E)$ = probability of x events of interest, given knowledge of n , N , and E

n = sample size

N = population size

E = number of events of interest in the population

$N - E$ = number of events that are not of interest in the population

x = number of events of interest in the sample

$$\binom{E}{x} = {}_E C_x = \text{number of combinations [see Equation (5.10) on page 196]}$$

$$x \leq E$$

$$x \leq n$$

Because the number of events of interest in the sample, represented by x , cannot be greater than the number of events of interest in the population, E , nor can x be greater than the sample size, n , the range of the hypergeometric random variable is limited to the sample size or to the number of events of interest in the population, whichever is smaller.

Equation (5.16) defines the mean of the hypergeometric distribution, and Equation (5.17) defines the standard deviation.

MEAN OF THE HYPERGEOMETRIC DISTRIBUTION

$$\mu = E(X) = \frac{nE}{N} \quad (5.16)$$

STANDARD DEVIATION OF THE HYPERGEOMETRIC DISTRIBUTION

$$\sigma = \sqrt{\frac{nE(N - E)}{N^2}} \sqrt{\frac{N - n}{N - 1}} \quad (5.17)$$

In Equation (5.17), the expression $\sqrt{\frac{N - n}{N - 1}}$ is a **finite population correction factor** that results from sampling without replacement from a finite population.

To illustrate the hypergeometric distribution, suppose that you are forming a team of 8 managers from different departments within your company. Your company has a total of 30 managers, and 10 of these managers are from the finance department. If you are to randomly select members of the team, what is the probability that the team will contain 2 managers from the finance department? Here, the population of $N = 30$ managers within the company is finite. In addition, $E = 10$ are from the finance department. A team of $n = 8$ members is to be selected.

Using Equation (5.15),

$$\begin{aligned} P(x = 2 | n = 8, N = 30, E = 10) &= \frac{\binom{10}{2} \binom{20}{6}}{\binom{30}{8}} \\ &= \frac{\left(\frac{10!}{2!(8)!}\right) \left(\frac{(20)!}{(6)!(14)!}\right)}{\left(\frac{30!}{8!(22)!}\right)} \\ &= 0.298 \end{aligned}$$

Thus, the probability that the team will contain two members from the finance department is 0.298, or 29.8%.

Computing hypergeometric probabilities can be tedious, especially as N gets large. Figure 5.5 shows how Excel and Minitab compute hypergeometric probabilities for the team formation example.

FIGURE 5.5

Excel and Minitab results for computing hypergeometric probabilities for the team formation problem

A		B	Probability Density Function
1 Hypergeometric Probabilities			Hypergeometric with N = 30, M = 10, and n = 8
2			
3 Data			
4 Sample size		8	x P(X = x)
5 No. of events of interest in population		10	0 0.021523
6 Population size		30	1 0.132447
7			2 0.298005
8 Hypergeometric Probabilities Table			3 0.317872
9	X	P(X)	4 0.173836
10	0	0.0215	5 0.049083
11	1	=HYPGEOM.DIST(A10, \$B\$4, \$B\$5, \$B\$6, FALSE)	6 0.006817
12	2	=HYPGEOM.DIST(A11, \$B\$4, \$B\$5, \$B\$6, FALSE)	7 0.000410
13	3	=HYPGEOM.DIST(A12, \$B\$4, \$B\$5, \$B\$6, FALSE)	8 0.000008
14	4	=HYPGEOM.DIST(A13, \$B\$4, \$B\$5, \$B\$6, FALSE)	
15	5	=HYPGEOM.DIST(A14, \$B\$4, \$B\$5, \$B\$6, FALSE)	
16	6	=HYPGEOM.DIST(A15, \$B\$4, \$B\$5, \$B\$6, FALSE)	
17	7	=HYPGEOM.DIST(A16, \$B\$4, \$B\$5, \$B\$6, FALSE)	
18	8	=HYPGEOM.DIST(A17, \$B\$4, \$B\$5, \$B\$6, FALSE)	

Example 5.7 shows an application of the hypergeometric distribution in portfolio selection.

EXAMPLE 5.7

Computing Hypergeometric Probabilities

You are a financial analyst facing the task of selecting mutual funds to purchase for a client's portfolio. You have narrowed the funds to be selected to 10 different funds. In order to diversify your client's portfolio, you will recommend the purchase of 4 different funds. Six of the funds are growth funds. What is the probability that of the 4 funds selected, 3 are growth funds?

SOLUTION Using Equation (5.15) with $X = 3, n = 4, N = 10$, and $E = 6$,

$$\begin{aligned}
 P(X = 3 | n = 4, N = 10, E = 6) &= \frac{\binom{6}{3} \binom{4}{1}}{\binom{10}{4}} \\
 &= \frac{\left(\frac{6!}{3!(3)!}\right) \left(\frac{(4)!}{(1)!(3)!}\right)}{\left(\frac{10!}{4!(6)!}\right)} \\
 &= 0.3810
 \end{aligned}$$

The probability that of the 4 funds selected, 3 are growth funds, is 0.3810, or 38.10%.

Problems for Section 5.5

LEARNING THE BASICS

5.42 Determine the following:

- a. If $n = 4$, $N = 10$, and $E = 5$, find $P(X = 3)$.
 - b. If $n = 4$, $N = 6$, and $E = 3$, find $P(X = 1)$.
 - c. If $n = 5$, $N = 12$, and $E = 3$, find $P(X = 0)$.
 - d. If $n = 3$, $N = 10$, and $E = 3$, find $P(X = 3)$.

5.43 Referring to Problem 5.42, compute the mean and standard deviation for the hypergeometric distributions described in (a) through (d).

APPLYING THE CONCEPTS

 **SELF TEST** **5.44** An auditor for the Internal Revenue Service is selecting a sample of 6 tax returns for an audit. If 2 or more of these returns are “improper,” the entire population of 100 tax returns will be audited. What is the probability that the entire population will be audited if the true number of improper returns in the population is

- a. 25?
 - b. 30?
 - c. 5?
 - d. 10?

e. Discuss the differences in your results, depending on the true number of improper returns in the population.

5.45 KSDLDS-Pros, an IT project management consulting firm, is forming an IT project management team of 5 professionals. In the firm of 50 professionals, 8 are considered to be data analytics specialists. If the professionals are selected at random, what is the probability that the team will include

- probability that the team will include

 - a. no data analytics specialist?
 - b. at least one data analytics specialist?

- c. no more than two data analytics specialists?
 - d. What is your answer to (a) if the team consists of 7 members?

5.46 From an inventory of 30 cars being shipped to a local automobile dealer, 4 are SUVs. What is the probability that if 4 cars arrive at a particular dealership,

- a. all 4 are SUVs?
 - b. none are SUVs?
 - c. at least 1 is an SUV?
 - d. What are your answers to (a) through (c) if 6 cars being shipped are SUVs?

5.47 As a quality control manager, you are responsible for checking the quality level of AC adapters for tablet PCs that your company manufactures. You must reject a shipment if you find 4 defective units. Suppose a shipment of 40 AC adapters has 8 defective units and 32 nondefective units. If you sample 12 AC adapters, what's the probability that

- a. there will be no defective units in the shipment?
 - b. there will be at least 1 defective unit in the shipment?
 - c. there will be 4 defective units in the shipment?
 - d. the shipment will be accepted?

5.48 In Example 5.7 above, a financial analyst was facing the task of selecting mutual funds to purchase for a client's portfolio. Suppose that the number of funds had been narrowed to 12 funds instead of the 10 funds (still with 6 growth funds) in Example 5.7. What is the probability that of the 4 funds selected,

- a. exactly 1 is a growth fund?
 - b. at least 1 is a growth fund?
 - c. 3 are growth fund?
 - d. Compare the result of (c) to the result of Example 5.7.

5.6 Using the Poisson Distribution to Approximate the Binomial Distribution

You can use the Poisson distribution to approximate the binomial distribution when n is large and π is very small. The approximation gets better as n gets larger and π gets smaller. Read more about how to use this approximation in the [Section 5.6 online topic](#).

USING STATISTICS

Events of Interest at Ricknel Home Centers, Revisited

In the Ricknel Home Improvement scenario at the beginning of this chapter, you were an accountant for the Ricknel Home Improvement Company. The company's accounting information system automatically reviews order forms from online customers for possible mistakes. Any questionable invoices are tagged and included in a daily exceptions report. Knowing that the probability that an order will be tagged is 0.10, you were able to use the binomial distribution to determine the chance of finding a certain number of tagged forms in a sample of size four. There was a 65.6% chance that none of the forms would be tagged, a 29.2% chance that one would be tagged, and a 5.2% chance

that two or more would be tagged. You were also able to determine that, on average, you would expect 0.4 form to be tagged, and the standard deviation of the number of tagged order forms would be 0.6. Now that you have learned the mechanics of using the binomial distribution for a known probability of 0.10 and a sample size of four, you will be able to apply the same approach to any given probability and sample size. Thus, you will be able to make inferences about the online ordering process and, more importantly, evaluate any changes or proposed changes to the process.



Sebastian Kaulitzki/Shutterstock

SUMMARY

In this chapter, you have studied the probability distribution for a discrete variable, the covariance and its application in finance, and three important discrete probability distributions: the binomial, Poisson, and hypergeometric distributions. In the next chapter, you will study several important continuous distributions, including the normal distribution.

To help decide which discrete probability distribution to use for a particular situation, you need to ask the following questions:

- Is there a fixed number of observations, n , each of which is classified as an event of interest or not an

event of interest? Is there an area of opportunity? If there is a fixed number of observations, n , each of which is classified as an event of interest or not an event of interest, you use the binomial or hypergeometric distribution. If there is an area of opportunity, you use the Poisson distribution.

- In deciding whether to use the binomial or hypergeometric distribution, is the probability of an event of interest constant over all trials? If yes, you can use the binomial distribution. If no, you can use the hypergeometric distribution.

REFERENCES

1. Bernstein, P. L. *Against the Gods: The Remarkable Story of Risk*. New York: Wiley, 1996.
2. Emery, D. R., J. D. Finnerty, and J. D. Stowe. *Corporate Financial Management*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2007.
3. Levine, D. M., P. Ramsey, and R. Smidt. *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.
4. Microsoft Excel 2013. Redmond, WA: Microsoft Corp., 2012.
5. Minitab Release 16. State College, PA: Minitab, Inc., 2010.
6. Taleb, N. *The Black Swan*, 2nd ed. New York: Random House, 2010.

KEY EQUATIONS

Expected Value, μ , of a Discrete Variable

$$\mu = E(X) = \sum_{i=1}^N x_i P(X = x_i) \quad (5.1)$$

Variance of a Discrete Variable

$$\sigma^2 = \sum_{i=1}^N [x_i - E(X)]^2 P(X = x_i) \quad (5.2)$$

Standard Deviation of a Discrete Variable

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^N x_i [x_i - E(X)]^2 P(X = x_i)} \quad (5.3)$$

Covariance

$$\sigma_{XY} = \sum_{i=1}^N [x_i - E(X)][y_i - E(Y)] P(x_i, y_i) \quad (5.4)$$

Expected Value of the Sum of Two Variables

$$E(X + Y) = E(X) + E(Y) \quad (5.5)$$

Variance of the Sum of Two Variables

$$Var(X + Y) = \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} \quad (5.6)$$

Standard Deviation of the Sum of Two Variables

$$\sigma_{X+Y} = \sqrt{\sigma_{X+Y}^2} \quad (5.7)$$

Portfolio Expected Return

$$E(P) = wE(X) + (1 - w)E(Y) \quad (5.8)$$

Portfolio Risk

$$\sigma_p = \sqrt{w^2\sigma_X^2 + (1 - w)^2\sigma_Y^2 + 2w(1 - w)\sigma_{XY}} \quad (5.9)$$

Combinations

$${}_nC_x = \frac{n!}{x!(n - x)!} \quad (5.10)$$

Binomial Distribution

$$P(X = x | n, \pi) = \frac{n!}{x!(n - x)!} \pi^x (1 - \pi)^{n-x} \quad (5.11)$$

Mean of the Binomial Distribution

$$\mu = E(X) = n\pi \quad (5.12)$$

Standard Deviation of the Binomial Distribution

$$\sigma = \sqrt{\sigma^2} = \sqrt{Var(X)} = \sqrt{n\pi(1 - \pi)} \quad (5.13)$$

Poisson Distribution

$$P(X = x | \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad (5.14)$$

Hypergeometric Distribution

$$P(X = x | n, N, E) = \frac{\binom{E}{x} \binom{N-E}{n-x}}{\binom{N}{n}} \quad (5.15)$$

Mean of the Hypergeometric Distribution

$$\mu = E(X) = \frac{nE}{N} \quad (5.16)$$

Standard Deviation of the Hypergeometric Distribution

$$\sigma = \sqrt{\frac{nE(N - E)}{N^2}} \sqrt{\frac{N-n}{N-1}} \quad (5.17)$$

KEY TERMS

area of opportunity 202
 binomial distribution 195
 covariance of a probability distribution (σ_{XY}) 190
 expected value 186
 expected value of the sum of two variables 191
 finite population correction factor 207
 hypergeometric distribution 206

mathematical model 195
 Poisson distribution 202
 portfolios 191
 portfolio expected return 191
 portfolio risk 191
 probability distribution for a discrete variable 186
 probability distribution function 195
 rule of combinations 196

standard deviation of a discrete variable 187
 standard deviation of the sum of two variables 191
 variance of a discrete variable 187
 variance of the sum of two variables 191

CHECKING YOUR UNDERSTANDING

5.49 What is the meaning of the expected value of a random variable?

5.50 What are the four properties that must be present in order to use the binomial distribution?

5.51 What are the four properties that must be present in order to use the Poisson distribution?

5.52 When do you use the hypergeometric distribution instead of the binomial distribution?

CHAPTER REVIEW PROBLEMS

5.53 Darwin Head, a 35-year-old sawmill worker, won \$1 million and a Chevrolet Malibu Hybrid by scoring 15 goals within 24 seconds at the Vancouver Canucks National Hockey League game (B. Ziemer, “Darwin Evolves into an Instant Millionaire,” *Vancouver Sun*, February 28, 2008, p. 1). Head said he would use the money to pay off his mortgage and provide for his children, and he had no plans to quit his job. The contest was part of the Chevrolet Malibu Million Dollar Shootout, sponsored by General Motors Canadian Division. Did GM-Canada risk the \$1 million? No! GM-Canada purchased event insurance from a company specializing in promotions at sporting events such as a half-court basketball shot or a hole-in-one giveaway at the local charity golf outing. The event insurance company estimates the probability of a contestant winning the contest, and for a modest charge, insures the event. The promoters pay the insurance premium but take on no added risk as the insurance company will make the large payout in the unlikely event that a contestant wins. To see how it works, suppose that the insurance company estimates that the probability a contestant would win a million-dollar shootout is 0.001 and that the insurance company charges \$4,000.

- Calculate the expected value of the profit made by the insurance company.
- Many call this kind of situation a win-win opportunity for the insurance company and the promoter. Do you agree? Explain.

5.54 Between 1896—when the Dow Jones index was created—and 2012, the index rose in 65% of the years. (Sources: M. Hulbert, “What the Past Can’t Tell Investors,” *The New York Times*, January 3, 2010, p. BU2 and bit.ly/100zwvT.) Based on this information, and assuming a binomial distribution, what do you think is the probability that the stock market will rise

- next year?
- the year after next?
- in four of the next five years?
- in none of the next five years?
- For this situation, what assumption of the binomial distribution might not be valid?

5.55 Smartphone adoption among American teens has increased substantially, and mobile access to the Internet is pervasive. One in four teenagers are “cell mostly” Internet users—that is, they *mostly go online using their phone* and not using some other device such as a desktop or laptop computer. (Source: *Teens and Technology 2013*, Pew Research Center, bit.ly/101ciF1.)

If a sample of 10 American teens is selected, what is the probability that

- 4 are “cell mostly” Internet users?
- at least 4 are “cell mostly” Internet users?

- at most 8 are “cell mostly” Internet users?
- If you selected the sample in a particular geographical area and found that none of the 10 respondents are “cell mostly” Internet users, what conclusions might you reach about whether the percentage of “cell mostly” Internet users in this area was 25%?

5.56 One theory concerning the Dow Jones Industrial Average is that it is likely to increase during U.S. presidential election years. From 1964 through 2012, the Dow Jones Industrial Average increased in 10 of the 13 U.S. presidential election years. Assuming that this indicator is a random event with no predictive value, you would expect that the indicator would be correct 50% of the time.

- What is the probability of the Dow Jones Industrial Average increasing in 10 or more of the 13 U.S. presidential election years if the probability of an increase in the Dow Jones Industrial Average is 0.50?
- What is the probability that the Dow Jones Industrial Average will increase in 10 or more of the 13 U.S. presidential election years if the probability of an increase in the Dow Jones Industrial Average in any year is 0.75?

5.57 Medical billing errors and fraud are on the rise. According to Medical Billing Advocates of America, 8 out of 10 times, the medical bills that you get are not right. (Data extracted from “Services Diagnose, Treat Medical Billing Errors,” *USA Today*, June 20, 2012.) If a sample of 10 medical bills is selected, what is the probability that

- 0 medical bills will contain errors?
- exactly 5 medical bills will contain errors?
- more than 5 medical bills will contain errors?
- What are the mean and standard deviation of the probability distribution?

5.58 Refer to Problem 5.57. Suppose that a quality improvement initiative has reduced the percentage of medical bills containing errors to 40%. If a sample of 10 medical bills is selected, what is the probability that

- 0 medical bills will contain errors?
- exactly 5 medical bills will contain errors?
- more than 5 medical bills contain errors?
- What are the mean and standard deviation of the probability distribution?
- Compare the results of (a) through (c) to those of Problem 5.57 (a) through (c).

5.59 Social log-ins involve recommending or sharing an article that you read online. According to Janrain, in the first quarter of 2013, 46% signed in via Facebook compared with 34% for Google.

(Source: “Social Login Trends Across the Web for Q1 2013,” bit.ly/ZQCRSF.) If a sample of 10 social log-ins is selected, what is the probability that

- more than 5 signed in using Facebook?
- more than 5 signed in using Google?
- none signed in using Facebook?
- What assumptions did you have to make to answer (a) through (c)?

5.60 The Consumer Financial Protection Bureau’s consumer response team hears directly from consumers about the challenges they face in the marketplace, brings their concerns to the attention of financial institutions, and assists in addressing their complaints. Consumer response accepts complaints related to mortgages, bank accounts and services, private student loans, other consumer loans, and credit reporting. Of the consumers who registered a bank account and service complaint, 41% cited “account management” as the type of complaint; these complaints are related to opening, closing, or managing the account and address issues, such as confusing marketing, denial, fees, statements, and joint accounts. (Source: *Consumer Response: A Snapshot of Complaints Received*, 1.usa.gov/WZ9N8Q.) Consider a sample of 20 consumers who registered bank account and service complaints. Use the binomial model to answer the following questions:

- What is the expected value, or mean, of the binomial distribution?
- What is the standard deviation of the binomial distribution?
- What is the probability that 10 of the 20 consumers cited “account management” as the type of complaint?
- What is the probability that no more than 5 of the consumers cited “account management” as the type of complaint?
- What is the probability that 5 or more of the consumers cited “account management” as the type of complaint?

5.61 Refer to Problem 5.60. In the same time period, 27% of the consumers registering a bank account and service compliant cited “deposit and withdrawal” as the type of complaint; these are issues such as transaction holds and unauthorized transactions.

- What is the expected value, or mean, of the binomial distribution?
- What is the standard deviation of the binomial distribution?
- What is the probability that none of the 20 consumers cited “deposit and withdrawal” as the type of complaint?
- What is the probability that no more than 2 of the consumers cited “deposit and withdrawal” as the type of complaint?
- What is the probability that 3 or more of the consumers cited “deposit and withdrawal” as the type of complaint?

5.62 One theory concerning the S&P 500 Index is that if it increases during the first five trading days of the year, it is likely to increase during the entire year. From 1950 through 2012, the S&P 500 Index had these early gains in 40 years (in 2011 there was virtually no change). In 35 of these 40 years, the S&P 500 Index increased for the entire year. Assuming that this indicator is a random event with no predictive value, you would expect that the indicator would be correct 50% of the time. What is the probability of the S&P 500 Index increasing in 35 or more years if the true probability of an increase in the S&P 500 Index is

- 0.50?
- 0.70?
- 0.90?
- Based on the results of (a) through (c), what do you think is the probability that the S&P 500 Index will increase if there is an early gain in the first five trading days of the year? Explain.

5.63 *Spurious correlation* refers to the apparent relationship between variables that either have no true relationship or are related to other variables that have not been measured. One widely publicized stock market indicator in the United States that is an example of spurious correlation is the relationship between the winner of the National Football League Super Bowl and the performance of the Dow Jones Industrial Average in that year. The “indicator” states that when a team that existed before the National Football League merged with the American Football League wins the Super Bowl, the Dow Jones Industrial Average will increase in that year. (Of course, any correlation between these is spurious as one thing has absolutely nothing to do with the other!) Since the first Super Bowl was held in 1967 through 2012, the indicator has been correct 37 out of 46 times. Assuming that this indicator is a random event with no predictive value, you would expect that the indicator would be correct 50% of the time.

- What is the probability that the indicator would be correct 37 or more times in 46 years?
- What does this tell you about the usefulness of this indicator?

5.64 The National Insurance Crime Bureau says that Miami-Dade, Broward, and Palm Beach counties account for a substantial number of questionable insurance claims referred to investigators. (Source: “United Auto Courts Reports,” bit.ly/100DZi9.) Assume that the number of questionable insurance claims referred to investigators by Miami-Dade, Broward, and Palm Beach counties is distributed as a Poisson random variable with a mean of 10 per day.

- What assumptions need to be made so that the number of questionable insurance claims referred to investigators by Miami-Dade, Broward, and Palm Beach counties is distributed as a Poisson random variable?

Making the assumptions given in (a), what is the probability that

- 5 questionable insurance claims will be referred to investigators by Miami-Dade, Broward, and Palm Beach counties in a day?
- 10 or fewer questionable insurance claims will be referred to investigators by Miami-Dade, Broward, and Palm Beach counties in a day?
- 11 or more questionable insurance claims will be referred to investigators by Miami-Dade, Broward, and Palm Beach counties in a day?

5.65 In the Florida lottery Lotto game, you select six numbers from a pool of numbers from 1 to 53 (see flalottery.com). Each wager costs \$1. You win the jackpot if you match all six numbers that you have selected.

Find the probability of

- winning the jackpot.
- matching five numbers.
- matching four numbers.
- matching three numbers.
- matching two numbers.
- matching one number.
- matching none of the numbers.
- If you match zero, one, or two numbers, you do not win anything. What is the probability that you will not win anything?
- The Lotto ticket gives complete game rules and probabilities of matching zero through six numbers. The lottery ticket has the saying “A Win for Education” on the back of the ticket. Do you think Florida’s slogan and the printed complete game rules and probabilities of matching zero through six numbers is an ethical approach to running the lottery game?

CASES FOR CHAPTER 5

Managing Ashland MultiComm Services

The Ashland MultiComm Services (AMS) marketing department wants to increase subscriptions for its *3-For-All* telephone, cable, and Internet combined service. AMS marketing has been conducting an aggressive direct-marketing campaign that includes postal and electronic mailings and telephone solicitations. Feedback from these efforts indicates that including premium channels in this combined service is a very important factor for both current and prospective subscribers. After several brainstorming sessions, the marketing department has decided to add premium cable channels as a no-cost benefit of subscribing to the *3-For-All* service.

The research director, Mona Fields, is planning to conduct a survey among prospective customers to determine how many premium channels need to be added to the *3-For-All* service in order to generate a subscription to the service. Based on past campaigns and on industry-wide data, she estimates the following:

Number of Free Premium Channels	Probability of Subscriptions
0	0.02
1	0.04
2	0.06
3	0.07
4	0.08
5	0.085

- If a sample of 50 prospective customers is selected and no free premium channels are included in the *3-For-All* service offer, given past results, what is the probability that
 - fewer than 3 customers will subscribe to the *3-For-All* service offer?
 - 0 customers or 1 customer will subscribe to the *3-For-All* service offer?
 - more than 4 customers will subscribe to the *3-For-All* service offer?

- Suppose that in the actual survey of 50 prospective customers, 4 customers subscribe to the *3-For-All* service offer. What does this tell you about the previous estimate of the proportion of customers who would subscribe to the *3-For-All* service offer?
- Instead of offering no premium free channels as in Problem 1, suppose that two free premium channels are included in the *3-For-All* service offer. Given past results, what is the probability that
 - fewer than 3 customers will subscribe to the *3-For-All* service offer?
 - 0 customers or 1 customer will subscribe to the *3-For-All* service offer?
 - more than 4 customers will subscribe to the *3-For-All* service offer?
- Compare the results of (a) through (c) to those of 1.
- Suppose that in the actual survey of 50 prospective customers, 6 customers subscribe to the *3-For-All* service offer. What does this tell you about the previous estimate of the proportion of customers who would subscribe to the *3-For-All* service offer?
- What do the results in (e) tell you about the effect of offering free premium channels on the likelihood of obtaining subscriptions to the *3-For-All* service?

- Suppose that additional surveys of 50 prospective customers were conducted in which the number of free premium channels was varied. The results were as follows:

Number of Free Premium Channels	Number of Subscriptions
1	5
3	6
4	6
5	7

How many free premium channels should the research director recommend for inclusion in the *3-For-All* service? Explain.

Digital Case

Apply your knowledge about expected value and the covariance in this continuing Digital Case from Chapters 3 and 4.

Open **BullsAndBears.pdf**, a marketing brochure from EndRun Financial Services. Read the claims and examine the supporting data. Then answer the following:

1. Are there any “catches” about the claims the brochure makes for the rate of return of Happy Bull and Worried Bear funds?

2. What subjective data influence the rate-of-return analyses of these funds? Could EndRun be accused of making false and misleading statements? Why or why not?
3. The expected-return analysis seems to show that the Worried Bear fund has a greater expected return than the Happy Bull fund. Should a rational investor never invest in the Happy Bull fund? Why or why not?

CHAPTER 5 EXCEL GUIDE

EG5.1 The PROBABILITY DISTRIBUTION for a DISCRETE VARIABLE

Key Technique Use the **SUMPRODUCT**(*cell range 1, cell range 2*) function (see Appendix F) to compute the expected value and variance.

Example Compute the expected value, variance, and standard deviation for the number of interruptions per day data of Table 5.1 on page 186.

In-Depth Excel Use the **Discrete Variable workbook** as a model.

For the example, open to the **DATA worksheet** of the **Discrete Variable workbook**. The worksheet already contains the entries needed to compute the expected value, variance, and standard deviation (shown in the **COMPUTE** worksheet) for the example.

For other problems, modify the **DATA** worksheet. Enter the probability distribution data into columns **A** and **B** and, if necessary, extend columns **C** through **E**, first selecting cell range **C7:E7** and then copying that cell range down as many rows as necessary. If the probability distribution has fewer than six outcomes, select the rows that contain the extra, unwanted outcomes, right-click, and then click **Delete** in the shortcut menu.

Read the **SHORT TAKES** for Chapter 5 for an explanation of the formulas found in the worksheets.

EG5.2 COVARIANCE of a PROBABILITY DISTRIBUTION and its APPLICATION in FINANCE

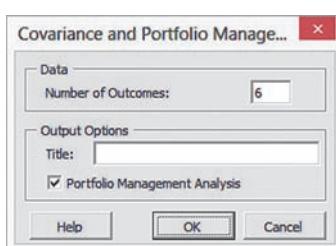
Key Technique Use the **SQRT** and **SUMPRODUCT** functions (see Appendix Section F) to help compute the portfolio analysis statistics.

Example Perform the portfolio analysis for the Section 5.2 investment example.

PHStat Use Covariance and Portfolio Analysis.

For the example, select **PHStat** → **Decision-Making** → **Covariance and Portfolio Analysis**. In the procedure's dialog box (shown below):

1. Enter **6** as the **Number of Outcomes**.
2. Enter a **Title**, check **Portfolio Management Analysis**, and click **OK**.



In the new worksheet (shown below):

1. Enter the probabilities and outcomes in the table that begins in cell **B3**.
2. Enter **0.5** as the **Weight assigned to X**.

In-Depth Excel Use the **COMPUTE worksheet** of the **Portfolio workbook** as a template.

The worksheet (shown below) already contains the data for the example. Overwrite the **X** and **P(X)** values and the weight assigned to the **X** value when you enter data for other problems. If a problem has more or fewer than three outcomes, first select row **5**, right-click, and click **Insert** (or **Delete**) in the shortcut menu to insert (or delete) rows one at a time. If you insert rows, select the cell range **B4:J4** and copy the contents of this range down through the new table rows.

A	B	C	D
1 Portfolio Expected Return and Risk			
3 Probabilities & Outcomes:	P	X	Y
4	0.2	-300	200
5	0.5	100	50
6	0.3	250	-100
7			
8 Weight Assigned to X	0.5		
9			
10 Statistics			
11 E(X)	65	=SUMPRODUCT(B4:B6, C4:C6)	
12 E(Y)	35	=SUMPRODUCT(B4:B6, D4:D6)	
13 Variance(X)	37525	=SUMPRODUCT(B4:B6, H4:H6)	
14 Standard Deviation(X)	193.7137	=SQRT(B13)	
15 Variance(Y)	11025	=SUMPRODUCT(B4:B6, I4:I6)	
16 Standard Deviation(Y)	105	=SQRT(B15)	
17 Covariance(XY)	-19275	=SUMPRODUCT(B4:B6, J4:J6)	
18 Variance(X+Y)	10000	=B13 + B15 + 2 * B17	
19 Standard Deviation(X+Y)	100	=SQRT(B18)	
20			
21 Portfolio Management			
22 Weight Assigned to X	0.5	=B8	
23 Weight Assigned to Y	0.5	=1 - B22	
24 Portfolio Expected Return	50	=B22 * B11 + B23 * B12	
25 Portfolio Risk	50	=SQRT(B22^2 * B13 + B23^2 * B15 + 2 * B22 * B23 * B17)	

The worksheet also contains a Calculations Area that contains various intermediate calculations. Open the **COMPUTE_FORMULAS worksheet** to examine all the formulas used in this area.

EG5.3 BINOMIAL DISTRIBUTION

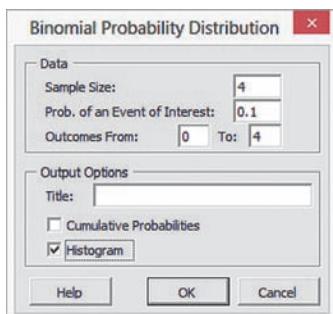
Key Technique Use the **BINOM.DIST**(*number of events of interest, sample size, probability of an event of interest, FALSE*) function.

Example Compute the binomial probabilities for $n = 4$ and $\pi = 0.1$, as is done in Figure 5.2 on page 199.

PHStat Use Binomial.

For the example, select **PHStat** → **Probability & Prob. Distributions** → **Binomial**. In the procedure's dialog box (shown on page 216):

1. Enter **4** as the **Sample Size**.
2. Enter **0.1** as the **Prob. of an Event of Interest**.
3. Enter **0** as the **Outcomes From** value and enter **4** as the **(Outcomes) To** value.
4. Enter a **Title**, check **Histogram**, and click **OK**.



Check **Cumulative Probabilities** before clicking **OK** in step 4 to have the procedure include columns for $P(\leq X)$, $P(<X)$, $P(>X)$, and $P(\geq X)$ in the binomial probabilities table.

In-Depth Excel Use the **Binomial workbook** as a template and model.

For the example, open to the **COMPUTE worksheet** of the **Binomial workbook**, shown in Figure 5.2 on page 199. The worksheet already contains the entries needed for the example. For other problems, change the sample size in cell **B4** and the probability of an event of interest in cell **B5**. If necessary, extend the binomial probabilities table by first selecting cell range **A18:B18** and then copying that cell range down as many rows as necessary. To construct a histogram of the probability distribution, use the Appendix Section B.9 instructions.

Read the **SHORT TAKES** for Chapter 5 for an explanation of the **CUMULATIVE** worksheet, which computes cumulative probabilities, and the worksheets to use with versions older than Excel 2010.

EG5.4 POISSON DISTRIBUTION

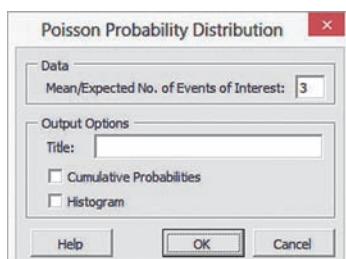
Key Technique Use the **POISSON.DIST(number of events of interest, the average or expected number of events of interest, FALSE)** function.

Example Compute the Poisson probabilities for the customer arrival problem in which $\lambda = 3$, as is done in Figure 5.4 on page 204.

PHStat Use **Poisson**.

For the example, select **PHStat → Probability & Prob. Distributions → Poisson**. In this procedure's dialog box (shown below):

1. Enter 3 as the **Mean/Expected No. of Events of Interest**.
2. Enter a Title and click **OK**.



Check **Cumulative Probabilities** before clicking **OK** in step 2 to have the procedure include columns for $P(\leq X)$, $P(<X)$, $P(>X)$,

and $P(\geq X)$ in the Poisson probabilities table. Check **Histogram** to construct a histogram of the Poisson probability distribution.

In-Depth Excel Use the **Poisson workbook** as a template.

For the example, open to the **COMPUTE worksheet** of the **Poisson workbook**, shown in Figure 5.4 on page 204. The worksheet already contains the entries for the example. For other problems, change the mean or expected number of events of interest in cell **E4**. To construct a histogram of the probability distribution, use the Appendix Section B.9 instructions.

Read the **SHORT TAKES** for Chapter 5 for an explanation of the **CUMULATIVE** worksheet, which computes cumulative probabilities, and the worksheets to use with versions older than Excel 2010.

EG5.5 HYPGEOMETRIC DISTRIBUTION

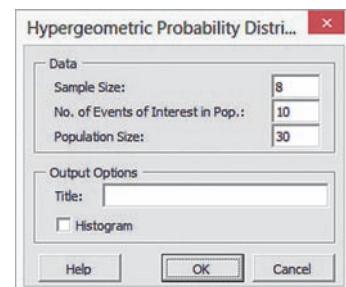
Key Technique Use the **HYPGEOM.DIST(X, sample size, number of events of interest in the population, population size, FALSE)** function.

Example Compute the hypergeometric probabilities for the team formation problem as is done in Figure 5.5 on page 207.

PHStat Use **Hypergeometric**.

For the example, select **PHStat → Probability & Prob. Distributions → Hypergeometric**. In this procedure's dialog box (shown below):

1. Enter 8 as the **Sample Size**.
2. Enter 10 as the **No. of Events of Interest in Pop.**.
3. Enter 30 as the **Population Size**.
4. Enter a **Title** and click **OK**.



Check **Histogram** to produce a histogram of the probability distribution.

In-Depth Excel Use the **Hypergeometric workbook** as a template.

For the example, open to the **COMPUTE worksheet** of the **Hypergeometric workbook**, shown in Figure 5.5 on page 207. The worksheet already contains the entries for the example. For other problems, change the sample size in cell **B4**, the number of events of interest in the population in cell **B5**, and the population size in cell **B6**. To construct a histogram of the probability distribution, use the Appendix Section B.9 instructions.

Read the **SHORT TAKES** for Chapter 5 for an explanation of the **CUMULATIVE** worksheet, which computes cumulative probabilities, and the worksheets to use with versions older than Excel 2010.

CHAPTER 5 MINITAB GUIDE

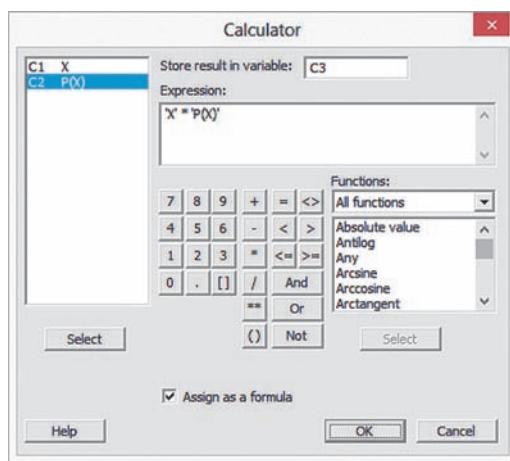
MG5.1 The PROBABILITY DISTRIBUTION for a DISCRETE VARIABLE

Expected Value of a Discrete Variable

Use **Calculator** to compute the expected value of a discrete random variable.

For example, to compute the expected value for the number of interruptions per day, open to the **Table_5.1 worksheet**. Select **Calc → Calculator**. In the Calculator dialog box (shown below):

1. Enter **C3** in the **Store result in variable** box and then press **Tab**. (C3 is the first empty column on the worksheet.)
2. Double-click **C1 X** in the variables list to add **X** to the **Expression** box.
3. Click ***** on the simulated keypad to add ***** to the **Expression** box.
4. Double-click **C2 P(X)** in the variables list to form the expression **X * 'P(X)'** in the **Expression** box.
5. Check **Assign as a formula**.
6. Click **OK**.



7. Enter **X*P(X)** as the name for **column C3**.

8. Reselect **Calc → Calculator**.

In the Calculator dialog box:

9. Enter **C4** in the **Store result in variable** box and then press **Tab**. (C4 is the first empty column on the worksheet.)
10. Enter **SUM(C3)** in the **Expression** box.
11. If necessary, clear **Assign as a formula**.
12. Click **OK**.

MG5.2 COVARIANCE and its APPLICATION in FINANCE

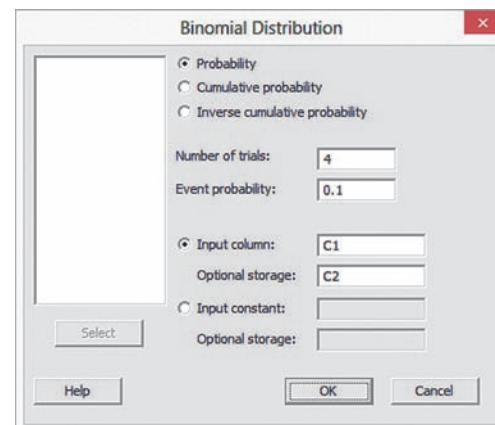
There are no Minitab instructions for this section.

MG5.3 BINOMIAL DISTRIBUTION

Use **Binomial** to compute binomial probabilities.

For example, to compute these probabilities for the Section 5.3 tagged orders example on page 196, open to a new, blank worksheet and:

1. Enter **X** as the name of **column C1**.
 2. Enter the values **0** through **4** in **column C1**, starting with row 1.
 3. Enter **P(X)** as the name of **column C2**.
 4. Select **Calc → Probability Distributions → Binomial**.
- In the Binomial Distribution dialog box (shown below):
5. Click **Probability** (to compute the probabilities of exactly **X** events of interest for all values of **X**).
 6. Enter **4** (the sample size) in the **Number of trials** box.
 7. Enter **0.1** in the **Event probability** box.
 8. Click **Input column**, enter **C1** in its box, and press **Tab**.
 9. Enter **C2** in the first **Optional storage** box.
 10. Click **OK**.



Skip step 9 to create the results shown in Figure 5.2 on page 199.

MG5.4 POISSON DISTRIBUTION

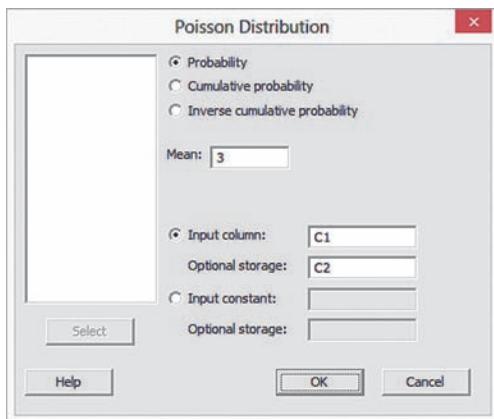
Use **Poisson** to compute Poisson probabilities.

For example, to compute these probabilities for the Section 5.4 bank customer arrivals example on page 203, open to a new, blank worksheet and:

1. Enter **X** as the name of **column C1**.
2. Enter the values **0** through **15** in **column C1**, starting with row 1.
3. Enter **P(X)** as the name of **column C2**.
4. Select **Calc → Probability Distributions → Poisson**.

In the Poisson Distribution dialog box (shown below):

1. Click **Probability** (to compute the probabilities of exactly X events of interest for all values of X).
2. Enter **3** in the **Mean** box.
3. Click **Input column**, enter **C1** in its box, and press **Tab**.
4. Enter **C2** in the first **Optional storage** box.
5. Click **OK**.



Skip step 8 to create the results shown in Figure 5.4 on page 204.

MG5.5 HYPERGEOMETRIC DISTRIBUTION

Use **Hypergeometric** to compute hypergeometric probabilities. For example, to compute these probabilities for the Section 5.5 team-formation example on page 207, open to a new, blank worksheet and:

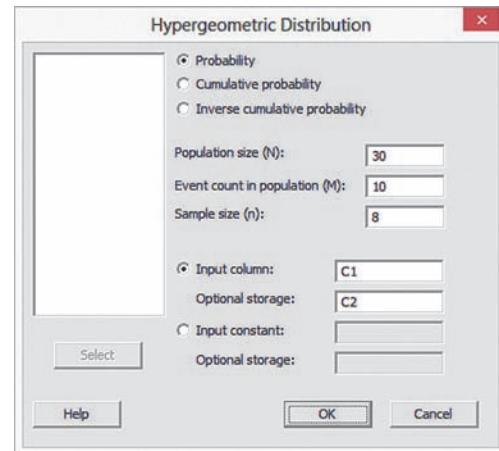
1. Enter **X** as the name of **column C1**.
2. Enter the values **0** through **8** in **column C1**, starting with row 1.

3. Enter **P(X)** as the name of **column C2**.

4. Select **Calc → Probability Distributions → Hypergeometric**.

In the Hypergeometric Distribution dialog box (shown below):

1. Click **Probability**.
2. Enter **30** in the **Population size (N)** box.
3. Enter **10** in the **Event count in population (M)** box.
4. Enter **8** in the **Sample size (n)** box.
5. Click **Input column**, enter **C1** in its box, and press **Tab**.
6. Enter **C2** in the first **Optional storage** box.
7. Click **OK**.



Skip step 6 to create the results shown in Figure 5.5 on page 207.

CHAPTER

6

The Normal Distribution and Other Continuous Distributions

CONTENTS

6.1 Continuous Probability Distributions

6.2 The Normal Distribution

VISUAL EXPLORATIONS:
Exploring the Normal Distribution

THINK ABOUT THIS: What Is Normal?

6.3 Evaluating Normality

6.4 The Uniform Distribution

6.5 The Exponential Distribution

6.6 The Normal Approximation to the Binomial Distribution (online)

USING STATISTICS: Normal Downloading at MyTVLab, Revisited

CHAPTER 6 EXCEL GUIDE

CHAPTER 6 MINITAB GUIDE

OBJECTIVES

To compute probabilities from the normal distribution

To use the normal distribution to solve business problems

To use the normal probability plot to determine whether a set of data is approximately normally distributed

To compute probabilities from the uniform distribution

To compute probabilities from the exponential distribution

USING STATISTICS

Normal Downloading at MyTVLab

You are a project manager for the MyTVLab website, an online service that streams movies and episodes from broadcast and cable TV series and that allows users to upload and share original videos. To attract and retain visitors to the website, you need to ensure that users can quickly download the exclusive-content daily videos.

To check how fast a video downloads, you open a web browser on a computer at the corporate offices of MyTVLab, load the MyTVLab home page, download the first website-exclusive video, and measure the download time. Download time—the amount of time in seconds, that passes from first clicking a download link until the video is ready to play—is a function of both the streaming media technology used and the number of simultaneous users of the website. Past data indicate that the mean download time is 7 seconds and that the standard deviation is 2 seconds. Approximately two-thirds of the download times are between 5 and 9 seconds, and about 95% of the download times are between 3 and 11 seconds. In other words, the download times are distributed as a bell-shaped curve, with a clustering around the mean of 7 seconds. How could you use this information to answer questions about the download times of the first video?



Cloki/Shutterstock

In Chapter 5, accounting managers at Ricknel Home Centers wanted to be able to answer questions about the number of tagged items in a given sample size. As a MyTVLab project manager, you face a different task—one that involves a continuous measurement because a download time could be any value and not just a whole number. How can you answer questions, such as the following, about this *continuous numerical variable*:

- What proportion of the video downloads take more than 9 seconds?
- How many seconds elapse before 10% of the downloads are complete?
- How many seconds elapse before 99% of the downloads are complete?
- How would enhancing the streaming media technology used affect the answers to these questions?

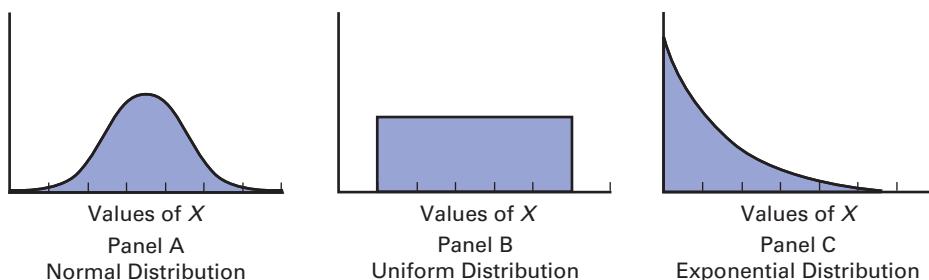
As in Chapter 5, you can use a probability distribution as a model. Reading this chapter will help you learn about characteristics of continuous probability distributions and how to use the normal distribution to solve business problems.

6.1 Continuous Probability Distributions

A **probability density function** is a mathematical expression that defines the distribution of the values for a continuous variable. Figure 6.1 graphically displays three probability density functions.

FIGURE 6.1

Three continuous probability distributions



Panel A depicts a *normal distribution*. The normal distribution is symmetrical and bell-shaped, implying that most observed values tend to cluster around the mean, which, due to the distribution's symmetrical shape, is equal to the median. Although the values in a normal distribution can range from negative infinity to positive infinity, the shape of the distribution makes it very unlikely that extremely large or extremely small values will occur.

Panel B shows a *uniform distribution* where the values are equally distributed in the range between the smallest value and the largest value. Sometimes referred to as the *rectangular distribution*, the uniform distribution is symmetrical, and therefore the mean equals the median.

Panel C illustrates an *exponential distribution*. This distribution is skewed to the right, making the mean larger than the median. The range for an exponential distribution is zero to positive infinity, but the distribution's shape makes it unlikely that extremely large values will occur.

6.2 The Normal Distribution

The **normal distribution** (also known as the *Gaussian distribution*) is the most common continuous distribution used in statistics. The normal distribution is vitally important in statistics for three main reasons:

- Numerous continuous variables common in business have distributions that closely resemble the normal distribution.
- The normal distribution can be used to approximate various discrete probability distributions.
- The normal distribution provides the basis for *classical statistical inference* because of its relationship to the *Central Limit Theorem* (which is discussed in Section 7.2).

The normal distribution is represented by the classic bell shape shown in Panel A of Figure 6.1. In the normal distribution, you can calculate the probability that values occur within certain ranges or intervals. However, because probability for continuous variables is measured

as an area under the curve, the probability of a *particular value* from a continuous distribution such as the normal distribution is zero. As an example, time (in seconds) is measured and not counted. Therefore, you can determine the probability that the download time for a video on a web browser is between 7 and 10 seconds, or the probability that the download time is between 8 and 9 seconds, or the probability that the download time is between 7.99 and 8.01 seconds. However, the probability that the download time is *exactly* 8 seconds is zero.

The normal distribution has several important theoretical properties:

- It is symmetrical, and its mean and median are therefore equal.
- It is bell-shaped in appearance.
- Its interquartile range is equal to 1.33 standard deviations. Thus, the middle 50% of the values are contained within an interval of two-thirds of a standard deviation below the mean and two-thirds of a standard deviation above the mean.
- It has an infinite range ($-\infty < X < \infty$).

In practice, many variables have distributions that closely resemble the theoretical properties of the normal distribution. The data in Table 6.1 represent the amount of soft drink in 10,000 1-liter bottles filled on a recent day. The continuous variable of interest, the amount of soft drink filled, can be approximated by the normal distribution. The measurements of the amount of soft drink in the 10,000 bottles cluster in the interval 1.05 to 1.055 liters and distribute symmetrically around that grouping, forming a bell-shaped pattern.

TABLE 6.1

Amount of Fill in
10,000 Bottles of a
Soft Drink

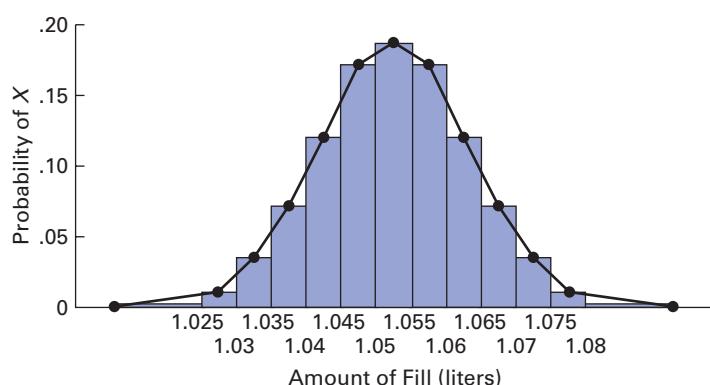
Amount of Fill (liters)	Relative Frequency
< 1.025	$48/10,000 = 0.0048$
$1.025 < 1.030$	$122/10,000 = 0.0122$
$1.030 < 1.035$	$325/10,000 = 0.0325$
$1.035 < 1.040$	$695/10,000 = 0.0695$
$1.040 < 1.045$	$1,198/10,000 = 0.1198$
$1.045 < 1.050$	$1,664/10,000 = 0.1664$
$1.050 < 1.055$	$1,896/10,000 = 0.1896$
$1.055 < 1.060$	$1,664/10,000 = 0.1664$
$1.060 < 1.065$	$1,198/10,000 = 0.1198$
$1.065 < 1.070$	$695/10,000 = 0.0695$
$1.070 < 1.075$	$325/10,000 = 0.0325$
$1.075 < 1.080$	$122/10,000 = 0.0122$
1.080 or above	$48/10,000 = 0.0048$
Total	1.0000

Figure 6.2 shows the relative frequency histogram and polygon for the distribution of the amount filled in 10,000 bottles.

FIGURE 6.2

Relative frequency histogram and polygon of the amount filled in 10,000 bottles of a soft drink

Source: Data are taken from Table 6.1.



For these data, the first three theoretical properties of the normal distribution are approximately satisfied. However, the fourth one, having an infinite range, is not. The amount filled in a bottle cannot possibly be zero or below, nor can a bottle be filled beyond its capacity. From Table 6.1, you see that only 48 out of every 10,000 bottles filled are expected to contain 1.08 liters or more, and an equal number are expected to contain less than 1.025 liters.

The symbol $f(X)$ is used to represent a probability density function. The **probability density function for the normal distribution** is given in Equation (6.1).

NORMAL PROBABILITY DENSITY FUNCTION

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)[(X-\mu)/\sigma]^2} \quad (6.1)$$

where

e = mathematical constant approximated by 2.71828

π = mathematical constant approximated by 3.14159

μ = mean

σ = standard deviation

X = any value of the continuous variable, where $-\infty < X < \infty$

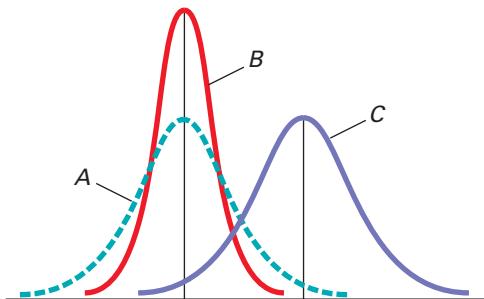
Student Tip

There is a different normal distribution for each combination of the mean, μ , and the standard deviation, σ .

Although Equation (6.1) may look complicated, the probabilities of the random variable X are dependent only on the mean, μ , and the standard deviation, σ , the two parameters of the normal distribution, because e and π are mathematical constants. There is a different normal distribution for each combination of the mean μ and the standard deviation σ . Figure 6.3 illustrates this principle. The distributions labeled A and B have the same mean (μ) but have different standard deviations. Distributions A and C have the same standard deviation (σ) but have different means. Distributions B and C have different values for both μ and σ .

FIGURE 6.3

Three normal distributions



Computing Normal Probabilities

To compute normal probabilities, you first convert a normally distributed variable, X , to a **standardized normal variable**, Z , using the **transformation formula**, shown in Equation (6.2). Applying this formula allows you to look up values in a normal probability table and avoid the tedious and complex computations that Equation (6.1) would otherwise require.

Z TRANSFORMATION FORMULA

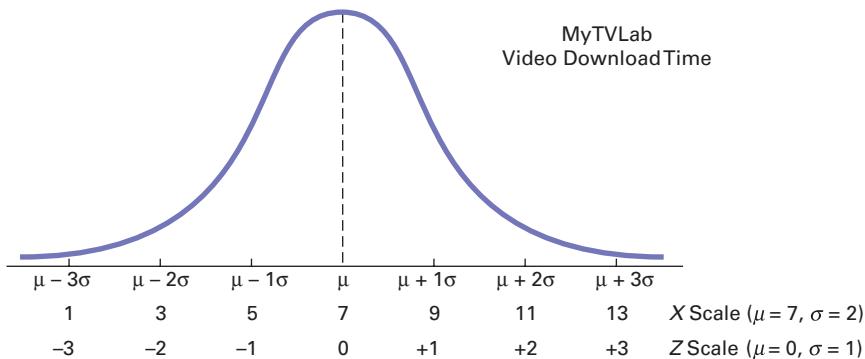
The Z value is equal to the difference between X and the mean, μ , divided by the standard deviation, σ .

$$Z = \frac{X - \mu}{\sigma} \quad (6.2)$$

The transformation formula computes a Z value that expresses the difference of the X value from the mean, μ , in standard deviation units (see Section 3.2 on page 113) called *standardized units*. While a variable, X , has mean, μ , and standard deviation, σ , the standardized variable, Z , always has mean $\mu = 0$ and standard deviation $\sigma = 1$.

Then you can determine the probabilities by using Table E.2, the **cumulative standardized normal distribution**. For example, recall from the Using Statistics scenario on page 219 that past data indicate that the time to download a video is normally distributed, with a mean $\mu = 7$ seconds and a standard deviation $\sigma = 2$ seconds. From Figure 6.4, you see that every measurement X has a corresponding standardized measurement Z , computed from Equation (6.2), the transformation formula.

FIGURE 6.4
Transformation of scales



Therefore, a download time of 9 seconds is equivalent to 1 standardized unit (1 standard deviation) above the mean because

$$Z = \frac{9 - 7}{2} = +1$$

A download time of 1 second is equivalent to -3 standardized units (3 standard deviations) below the mean because

$$Z = \frac{1 - 7}{2} = -3$$

In Figure 6.4, the standard deviation is the unit of measurement. In other words, a time of 9 seconds is 2 seconds (1 standard deviation) higher, or *slower*, than the mean time of 7 seconds. Similarly, a time of 1 second is 6 seconds (3 standard deviations) lower, or *faster*, than the mean time.

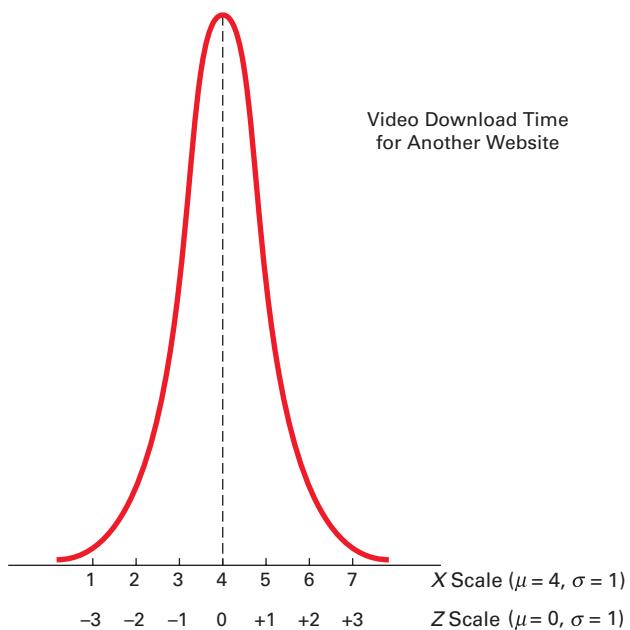
To further illustrate the transformation formula, suppose that another website has a download time for a video that is normally distributed, with a mean $\mu = 4$ seconds and a standard deviation $\sigma = 1$ second. Figure 6.5 on page 224 shows this distribution.

Comparing these results with those of the MyTVLab website, you see that a download time of 5 seconds is 1 standard deviation above the mean download time because

$$Z = \frac{5 - 4}{1} = +1$$

FIGURE 6.5

A different transformation of scales



A time of 1 second is 3 standard deviations below the mean download time because

$$Z = \frac{1 - 4}{1} = -3$$

With the Z value computed, you look up the normal probability using a table of values from the cumulative standardized normal distribution, such as Table E.2 in Appendix E. Suppose you wanted to find the probability that the download time for the MyTVLab website is less than 9 seconds. Recall from page 223 that transforming $X = 9$ to standardized Z units, given a mean $\mu = 7$ seconds and a standard deviation $\sigma = 2$ seconds, leads to a Z value of +1.00.

With this value, you use Table E.2 to find the cumulative area under the normal curve less than (to the left of) $Z = +1.00$. To read the probability or area under the curve less than $Z = +1.00$, you scan down the Z column in Table E.2 until you locate the Z value of interest (in 10ths) in the Z row for 1.0. Next, you read across this row until you intersect the column that contains the 100ths place of the Z value. Therefore, in the body of the table, the probability for $Z = 1.00$ corresponds to the intersection of the row $Z = 1.0$ with the column $Z = .00$. Table 6.2, which reproduces a portion of Table E.2, shows this intersection. The probability

Student Tip

Remember that when dealing with a continuous distribution such as the normal, the word *area* has the same meaning as *probability*.

TABLE 6.2

Finding a Cumulative Area Under the Normal Curve

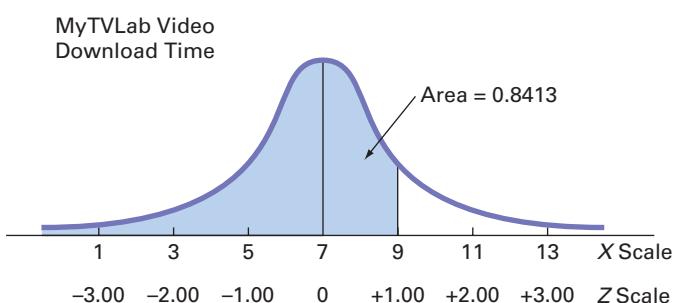
Z	Cumulative Probabilities									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7518	.7549
0.7	.7580	.7612	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621

Source: Extracted from Table E.2.

listed at the intersection is 0.8413, which means that there is an 84.13% chance that the download time will be less than 9 seconds. Figure 6.6 graphically shows this probability.

FIGURE 6.6

Determining the area less than Z from a cumulative standardized normal distribution



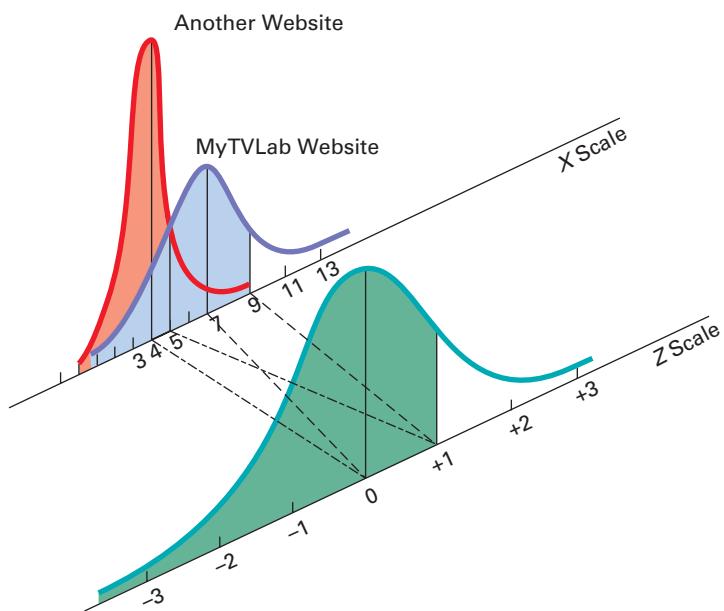
However, for the other website, you see that a time of 5 seconds is 1 standardized unit above the mean time of 4 seconds. Thus, the probability that the download time will be less than 5 seconds is also 0.8413. Figure 6.7 shows that regardless of the value of the mean, μ , and standard deviation, σ , of a normally distributed variable, Equation (6.2) can transform the X value to a Z value.

FIGURE 6.7

Demonstrating a transformation of scales for corresponding cumulative portions under two normal curves

Student Tip

You will find it very helpful when computing probabilities under the normal curve if you draw a normal curve and then enter the values for the mean and X below the curve and shade the desired area to be determined under the curve.



Now that you have learned to use Table E.2 with Equation (6.2), you can answer many questions related to the MyTVLab video download, using the normal distribution.

EXAMPLE 6.1

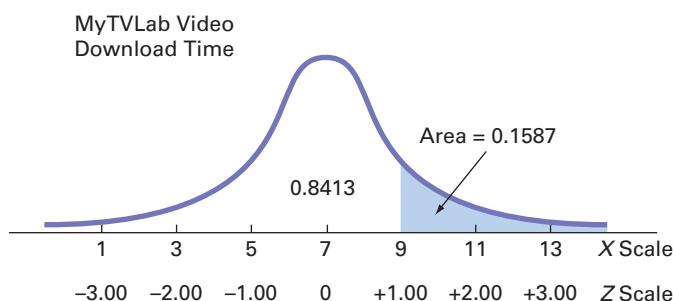
Finding $P(X > 9)$

What is the probability that the video download time for the MyTVLab website will be more than 9 seconds?

SOLUTION The probability that the download time will be less than 9 seconds is 0.8413 (see Figure 6.6 above). Thus, the probability that the download time will be more than 9 seconds is the complement of less than 9 seconds, $1 - 0.8413 = 0.1587$. Figure 6.8 illustrates this result.

FIGURE 6.8

Finding $P(X > 9)$



EXAMPLE 6.2

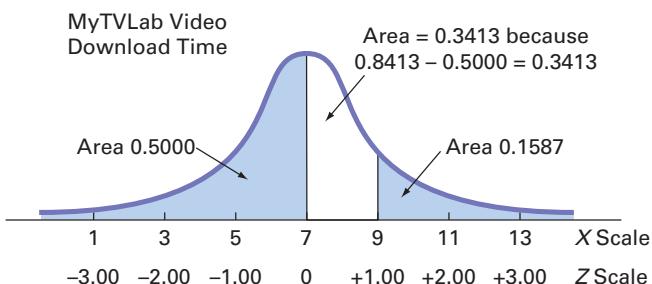
Finding $P(X < 7$ or $X > 9)$

What is the probability that the video download time for the MyTVLab website will be less than 7 seconds or more than 9 seconds?

SOLUTION To find this probability, you separately calculate the probability of a download time less than 7 seconds and the probability of a download time greater than 9 seconds and then add these two probabilities together. Figure 6.9 illustrates this result.

FIGURE 6.9

Finding $P(X < 7$ or $X > 9)$



Because the mean is 7 seconds, and because the mean is equal to the median in a normal distribution, 50% of download times are under 7 seconds. From Example 6.1, you know that the probability that the download time is greater than 9 seconds is 0.1587. Therefore, the probability that a download time is under 7 or over 9 seconds, $P(X < 7 \text{ or } X > 9)$, is $0.5000 + 0.1587 = 0.6587$.

EXAMPLE 6.3

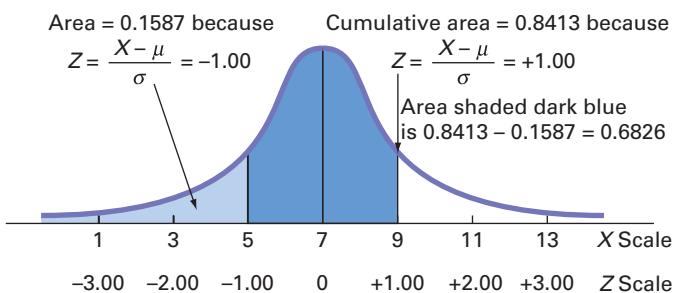
Finding $P(5 < X < 9)$

What is the probability that video download time for the MyTVLab website will be between 5 and 9 seconds—that is, $P(5 < X < 9)$?

SOLUTION In Figure 6.10, you can see that the area of interest is located between two values, 5 and 9.

FIGURE 6.10

Finding $P(5 < X < 9)$



In Example 6.1 on page 225, you already found that the area under the normal curve less than 9 seconds is 0.8413. To find the area under the normal curve less than 5 seconds,

$$Z = \frac{5 - 7}{2} = -1.00$$

Using Table E.2, you look up $Z = -1.00$ and find 0.1587. Therefore, the probability that the download time will be between 5 and 9 seconds is $0.8413 - 0.1587 = 0.6826$, as displayed in Figure 6.10.

The result of Example 6.3 enables you to state that for any normal distribution, 68.26% of the values are within ± 1 standard deviation of the mean. From Figure 6.11, you can see that 95.44% of the values are within ± 2 standard deviations of the mean. Thus, 95.44% of the download times are between 3 and 11 seconds. From Figure 6.12, you can see that 99.73% of the values are within ± 3 standard deviations above or below the mean.

FIGURE 6.11

Finding $P(3 < X < 11)$

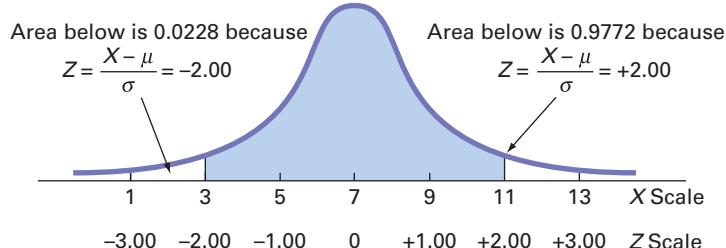
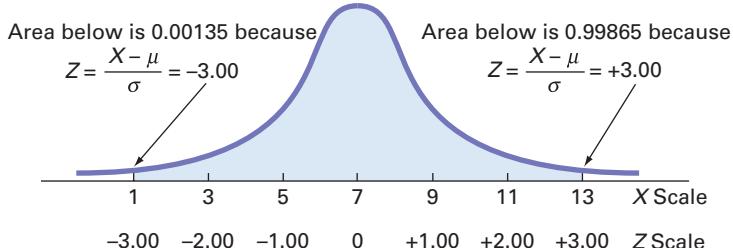


FIGURE 6.12

Finding $P(1 < X < 13)$



Thus, 99.73% of the download times are between 1 and 13 seconds. Therefore, it is unlikely (0.0027, or only 27 in 10,000) that a download time will be so fast or so slow that it will take under 1 second or more than 13 seconds. In general, you can use 6σ (i.e., 3 standard deviations below the mean to 3 standard deviations above the mean) as a practical approximation of the range for normally distributed data.

Figures 6.10, 6.11, and 6.12 illustrate that for any normal distribution,

- Approximately 68.26% of the values fall within ± 1 standard deviation of the mean
- Approximately 95.44% of the values fall within ± 2 standard deviations of the mean
- Approximately 99.73% of the values fall within ± 3 standard deviations of the mean

This result is the justification for the empirical rule presented on page 129. The accuracy of the empirical rule improves as a data set follows the normal distribution more closely.

Finding X Values

Examples 6.1 through 6.3 require you to use the normal distribution Table E.2 to find an area under the normal curve that corresponds to a specific X value. For other situations, you may need to do the reverse: Find the X value that corresponds to a specific area. In general, you use Equation (6.3) for finding an X value.

FINDING AN X VALUE ASSOCIATED WITH A KNOWN PROBABILITY

The X value is equal to the mean, μ , plus the product of the Z value and the standard deviation, σ .

$$X = \mu + Z\sigma \quad (6.3)$$

To find a *particular* value associated with a known probability, follow these steps:

- Sketch the normal curve and then place the values for the mean and X on the X and Z scales.
- Find the cumulative area less than X .
- Shade the area of interest.
- Using Table E.2, determine the Z value corresponding to the area under the normal curve less than X .
- Using Equation (6.3), solve for X :

$$X = \mu + Z\sigma$$

Examples 6.4 and 6.5 illustrate this technique.

EXAMPLE 6.4

Finding the X Value for a Cumulative Probability of 0.10

How much time (in seconds) will elapse before the fastest 10% of the downloads of a MyTVLab video are complete?

SOLUTION Because 10% of the videos are expected to download in under X seconds, the area under the normal curve less than this value is 0.1000. Using the body of Table E.2, you search for the area or probability of 0.1000. The closest result is 0.1003, as shown in Table 6.3 (which is extracted from Table E.2).

TABLE 6.3

Finding a Z Value Corresponding to a Particular Cumulative Area (0.10) Under the Normal Curve

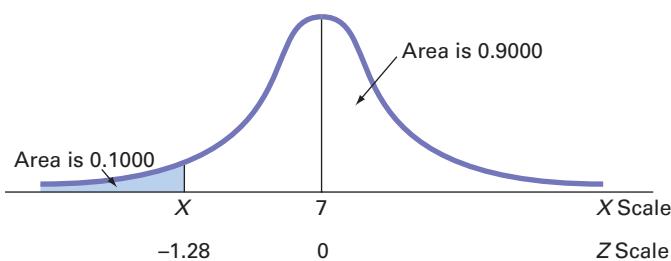
Z	Cumulative Probabilities									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985

Source: Extracted from Table E.2.

Working from this area to the margins of the table, you find that the Z value corresponding to the particular Z row (-1.2) and Z column (.08) is -1.28 (see Figure 6.13).

FIGURE 6.13

Finding Z to determine X



Once you find Z , you use Equation (6.3) on page 227 to determine the X value. Substituting $\mu = 7$, $\sigma = 2$, and $Z = -1.28$,

$$X = \mu + Z\sigma$$

$$X = 7 + (-1.28)(2) = 4.44 \text{ seconds}$$

Thus, 10% of the download times are 4.44 seconds or less.

EXAMPLE 6.5

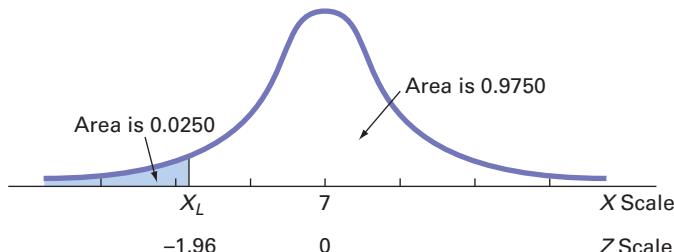
Finding the X Values That Include 95% of the Download Times

FIGURE 6.14

Finding Z to determine X_L

What are the lower and upper values of X , symmetrically distributed around the mean, that include 95% of the download times for a video at the MyTVLab website?

SOLUTION First, you need to find the lower value of X (called X_L). Then, you find the upper value of X (called X_U). Because 95% of the values are between X_L and X_U , and because X_L and X_U are equally distant from the mean, 2.5% of the values are below X_L (see Figure 6.14).



Although X_L is not known, you can find the corresponding Z value because the area under the normal curve less than this Z is 0.0250. Using the body of Table 6.4, you search for the probability 0.0250.

TABLE 6.4

Finding a Z Value Corresponding to a Cumulative Area of 0.025 Under the Normal Curve

Z	Cumulative Area									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294

Source: Extracted from Table E.2.

Working from the body of the table to the margins of the table, you see that the Z value corresponding to the particular Z row (-1.9) and Z column ($.06$) is -1.96 .

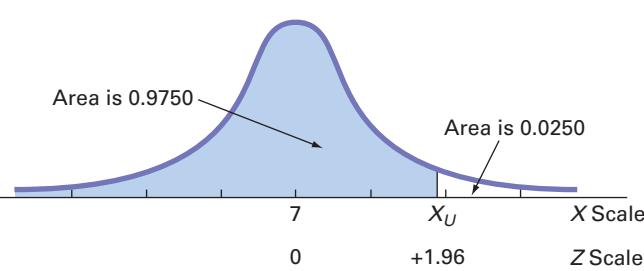
Once you find Z , the final step is to use Equation (6.3) on page 227 as follows:

$$\begin{aligned} X &= \mu + Z\sigma \\ &= 7 + (-1.96)(2) \\ &= 7 - 3.92 \\ &= 3.08 \text{ seconds} \end{aligned}$$

You use a similar process to find X_U . Because only 2.5% of the video downloads take longer than X_U seconds, 97.5% of the video downloads take less than X_U seconds. From the symmetry of the normal distribution, you find that the desired Z value, as shown in Figure 6.15, is $+1.96$ (because Z lies to the right of the standardized mean of 0). You can also extract this Z value from Table 6.5. You can see that 0.975 is the area under the normal curve less than the Z value of $+1.96$.

FIGURE 6.15

Finding Z to determine X_U



(continued)

TABLE 6.5

Finding a Z Value Corresponding to a Cumulative Area of 0.975 Under the Normal Curve

Cumulative Area										
Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
+1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
+1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
+2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

Source: Extracted from Table E.2.

Using Equation (6.3) on page 227,

$$\begin{aligned}
 X &= \mu + Z\sigma \\
 &= 7 + (+1.96)(2) \\
 &= 7 + 3.92 \\
 &= 10.92 \text{ seconds}
 \end{aligned}$$

Therefore, 95% of the download times are between 3.08 and 10.92 seconds.

Instead of looking up cumulative probabilities in a table, you can use Excel or Minitab to compute normal probabilities. Figure 6.16 displays an Excel worksheet that computes normal probabilities and finds X values for problems similar to Examples 6.1 through 6.5. Figure 6.17 shows Minitab results for Examples 6.1 and 6.4. (You need to subtract the results in the left part of the figure from 1.0 to obtain the answer to Example 6.1)

FIGURE 6.16

Excel worksheet for computing normal probabilities and finding X values (shown in two parts)

A	B
1 Normal Probabilities	
2	
3 Common Data	
4 Mean	7
5 Standard Deviation	2
6	
7 Probability for X <=	
8 X Value	7
9 Z Value	0
10 P(X<=7)	0.5000
	=STANDARDIZE(B8, B4, B5)
	=NORM.DIST(B8, B4, B5, TRUE)
11	
12 Probability for X >	
13 X Value	9
14 Z Value	1
15 P(X>9)	0.1587
	=1 - NORM.DIST(B13, B4, B5, TRUE)
16	
17 Probability for X<7 or X>9	
18 P(X<7 or X>9)	0.6587
	=B10 + B15
19	

A	B
20 Probability for a Range	
21 From X Value	5
22 To X Value	9
23 Z Value for 5	-1
24 Z Value for 9	1
25 P(X<=5)	0.1587
26 P(X<=9)	0.8413
27 P(5<=X<=9)	0.6827
	=ABS(B26 - B25)
28	
29 Find X and Z Given a Cum. Pctage.	
30 Cumulative Percentage	10.00%
31 Z Value	-1.28
32 X Value	4.44
33	
34 Find X Values Given a Percentage	
35 Percentage	95.00%
36 Z Value	-1.96
37 Lower X Value	3.08
38 Upper X Value	10.92
	=B4 + (B36 * B5)

FIGURE 6.17

Minitab results for Examples 6.1 and 6.4

Cumulative Distribution Function

Normal with mean = 7 and standard deviation = 2

x	P(X <= x)
9	0.841345

Inverse Cumulative Distribution Function

Normal with mean = 7 and standard deviation = 2

P(X <= x)	x
0.1	4.43690

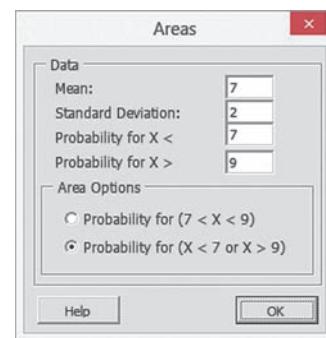
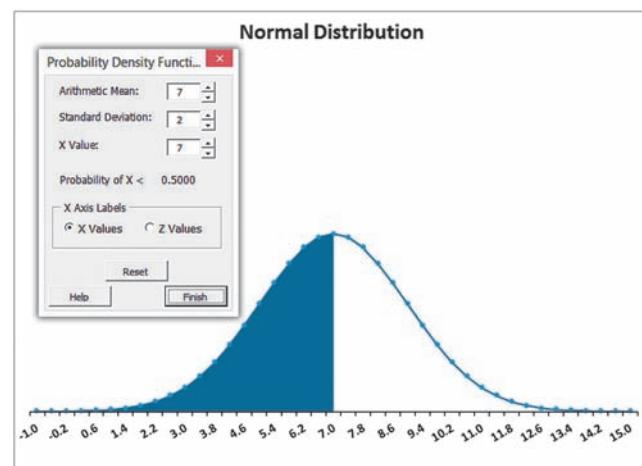
VISUAL EXPLORATIONS

Exploring the Normal Distribution

Open the **VE-Normal Distribution add-in workbook** to explore the normal distribution. (See Appendix C to learn how you can download a copy of this workbook and Appendix Section D.5 before using this workbook.) When this workbook opens properly, it adds a Normal Distribution menu in the Add-ins tab.

To explore the effects of changing the mean and standard deviation on the area under a normal distribution curve workbook, select **Add-ins → Normal Distribution → Probability Density Function**. The add-in displays a normal curve for the MyTVLab website download example and a floating control panel (shown at top right). Use the control panel spinner buttons to change the values for the mean, standard deviation, and X value and then note the effects of these changes on the probability of $X <$ value and the corresponding shaded area under the curve. To see the normal curve labeled with Z values, click **Z Values**. Click the **Reset** button to reset the control panel values. Click **Finish** to finish exploring.

To create shaded areas under the curve for problems similar to Examples 6.2 and 6.3, select **Add-ins → Normal Distribution → Areas**. In the Areas dialog box (shown at bottom right), enter values, select an Area Option, and click **OK**. The add-in creates a normal distribution curve with areas that are shaded according to the values you entered.



THINK ABOUT THIS

What Is Normal?

Ironically, the statistician who popularized the use of “normal” to describe the distribution discussed in Section 6.2 was someone who saw the distribution as anything but the everyday, anticipated occurrence that the adjective *normal* usually suggests.

Starting with an 1894 paper, Karl Pearson argued that measurements of phenomena do not naturally, or “normally,” conform to the classic bell shape. While this principle underlies much of statistics today, Pearson’s point of view was radical to contemporaries who saw the world as standardized and normal. Pearson changed minds by showing that some populations are naturally *skewed* (coining that term in passing), and he helped put to rest the notion that the normal distribution underlies all phenomena.

Today, people still make the type of mistake that Pearson refuted. As a student, you are probably familiar with discussions about grade inflation,

a real phenomenon at many schools. But have you ever realized that a “proof” of this inflation—that there are “too few” low grades because grades are skewed toward A’s and B’s—wrongly implies that grades should be “normally” distributed? Because college students represent small *nonrandom* samples, there are plenty of reasons to suspect that the distribution of grades would not be “normal.”

Misunderstandings about the normal distribution have occurred both in business and in the public sector through the years. These misunderstandings have caused a number of business blunders and have sparked several public policy debates, including the causes of the collapse of large financial institutions in 2008. According to one theory, the investment banking industry’s application of the normal distribution to assess risk may have contributed to the global collapse (see

“A Finer Formula for Assessing Risks,” *The New York Times*, May 11, 2010, p. B2 and reference 8). Using the normal distribution led these banks to overestimate the probability of having stable market conditions and underestimate the chance of unusually large market losses.

According to this theory, the use of other distributions that have less area in the middle of their curves, and, therefore, more in the “tails” that represent unusual market outcomes, may have led to less serious losses.

As you study this chapter, make sure you understand the assumptions that must hold for the proper use of the “normal” distribution, assumptions that were not explicitly verified by the investment bankers. And, most importantly, always remember that the name *normal distribution* does not mean normal in the everyday sense of the word.

Problems for Section 6.2

LEARNING THE BASICS

6.1 Given a standardized normal distribution (with a mean of 0 and a standard deviation of 1, as in Table E.2), what is the probability that

- a. Z is less than 1.57?
- b. Z is greater than 1.84?
- c. Z is between 1.57 and 1.84?
- d. Z is less than 1.57 or greater than 1.84?

6.2 Given a standardized normal distribution (with a mean of 0 and a standard deviation of 1, as in Table E.2), what is the probability that

- a. Z is between -1.57 and 1.84?
- b. Z is less than -1.57 or greater than 1.84?
- c. What is the value of Z if only 2.5% of all possible Z values are larger?
- d. Between what two values of Z (symmetrically distributed around the mean) will 68.26% of all possible Z values be contained?

6.3 Given a standardized normal distribution (with a mean of 0 and a standard deviation of 1, as in Table E.2), what is the probability that

- a. Z is less than 1.08?
- b. Z is greater than -0.21?
- c. Z is less than -0.21 or greater than the mean?
- d. Z is less than -0.21 or greater than 1.08?

6.4 Given a standardized normal distribution (with a mean of 0 and a standard deviation of 1, as in Table E.2), determine the following probabilities:

- a. $P(Z > 1.08)$
- b. $P(Z < -0.21)$
- c. $P(-1.96 < Z < -0.21)$
- d. What is the value of Z if only 15.87% of all possible Z values are larger?

6.5 Given a normal distribution with $\mu = 100$ and $\sigma = 10$, what is the probability that

- a. $X > 75$?
- b. $X < 70$?
- c. $X < 80$ or $X > 110$?
- d. Between what two X values (symmetrically distributed around the mean) are 80% of the values?

6.6 Given a normal distribution with $\mu = 50$ and $\sigma = 4$, what is the probability that

- a. $X > 43$?
- b. $X < 42$?
- c. 5% of the values are less than what X value?
- d. Between what two X values (symmetrically distributed around the mean) are 60% of the values?

APPLYING THE CONCEPTS

6.7 According to bottledwater.org, in 2012, the per capita consumption of bottled water in the United States was reported to be 30.8 gallons. Assume that the per capita consumption of bottled water in the United States is approximately normally distributed with a mean of 30.8 gallons and a standard deviation of 10 gallons.

- a. What is the probability that someone in the United States consumed more than 32 gallons of bottled water in 2012?
- b. What is the probability that someone in the United States consumed between 10 and 20 gallons of bottled water in 2012?

- c. What is the probability that someone in the United States consumed less than 10 gallons of bottled water in 2012?
- d. Ninety-nine percent of the people in the United States consumed less than how many gallons of bottled water?



6.8 Toby's Trucking Company determined that the distance traveled per truck per year is normally distributed, with a mean of 50 thousand miles and a standard deviation of 12 thousand miles.

- a. What proportion of trucks can be expected to travel between 34 and 50 thousand miles in a year?
- b. What percentage of trucks can be expected to travel either less than 30 or more than 60 thousand miles in a year?
- c. How many miles will be traveled by at least 80% of the trucks?
- d. What are your answers to (a) through (c) if the standard deviation is 10 thousand miles?

6.9 Consumers spend an average of \$13.80 on a meal at a restaurant in 2012. (Data extracted from www.alixpartners.com.) Assume that the amount spent on a restaurant meal is normally distributed and that the standard deviation is \$2.

- a. What is the probability that a randomly selected person spent more than \$15?
- b. What is the probability that a randomly selected person spent between \$10 and \$12?
- c. Between what two values will the middle 95% of the amounts spent fall?

6.10 A set of final examination grades in an introductory statistics course is normally distributed, with a mean of 73 and a standard deviation of 8.

- a. What is the probability that a student scored below 91 on this exam?
- b. What is the probability that a student scored between 65 and 89?
- c. The probability is 5% that a student taking the test scores higher than what grade?
- d. If the professor grades on a curve (i.e., gives A's to the top 10% of the class, regardless of the score), are you better off with a grade of 81 on this exam or a grade of 68 on a different exam, where the mean is 62 and the standard deviation is 3? Show your answer statistically and explain.

6.11 A Nielsen study indicates that mobile subscribers between 18 and 24 years of age spend a substantial amount of time watching video on their devices, reporting a mean of 396 minutes per month. (Data extracted from bit.ly/13f4uab.) Assume that the amount of time watching video on a mobile device per month is normally distributed and that the standard deviation is 50 minutes.

- a. What is the probability that an 18- to 24-year-old mobile subscriber spends less than 321 minutes watching video on his or her mobile device per month?
- b. What is the probability that an 18- to 24-year-old mobile subscriber spends between 320 minutes and 471 minutes watching video on his or her mobile device per month?
- c. What is the probability that an 18- to 24-year-old mobile subscriber spends more than 471 minutes watching video on his or her mobile device per month?
- d. One percent of all 18- to 24-year-old mobile subscribers will spend less than how many minutes watching video on his or her mobile device per month?

6.12 In 2012, the per capita consumption of soft drinks in the United States was reported to be 44 gallons. (Data extracted from [on-msn.com/XdwVIq](#).) Assume that the per capita consumption of soft drinks in the United States is approximately normally distributed with a mean of 44 gallons and a standard deviation of 14 gallons.

- What is the probability that someone in the United States consumed more than 60 gallons of soft drinks in 2012?
- What is the probability that someone in the United States consumed between 15 and 30 gallons of soft drinks in 2012?
- What is the probability that someone in the United States consumed less than 15 gallons of soft drinks in 2012?
- Ninety-nine percent of the people in the United States consumed less than how many gallons of soft drinks?

6.13 Many manufacturing problems involve the matching of machine parts, such as shafts that fit into a valve hole. A particular design requires a shaft with a diameter of 22.000 mm, but shafts with diameters between 21.990 mm and 22.010 mm are acceptable. Suppose that the manufacturing process yields shafts with diameters normally distributed, with a mean of 22.002 mm and a standard deviation of 0.005 mm. For this process, what is

- the proportion of shafts with a diameter between 21.99 mm and 22.00 mm?
- the probability that a shaft is acceptable?
- the diameter that will be exceeded by only 2% of the shafts?
- What would be your answers in (a) through (c) if the standard deviation of the shaft diameters were 0.004 mm?

6.3 Evaluating Normality

As first stated in Section 6.2, the normal distribution has several important theoretical properties:

- It is symmetrical; thus, the mean and median are equal.
- It is bell-shaped; thus, the empirical rule applies.
- The interquartile range equals 1.33 standard deviations.
- The range is approximately equal to 6 standard deviations.

As Section 6.2 notes, many continuous variables used in business closely follow a normal distribution. To determine whether a set of data can be approximated by the normal distribution, you either compare the characteristics of the data with the theoretical properties of the normal distribution or construct a normal probability plot.

Comparing Data Characteristics to Theoretical Properties

Many continuous variables have characteristics that approximate theoretical properties. However, other continuous variables are often neither normally distributed nor approximately normally distributed. For such variables, the descriptive characteristics of the data are inconsistent with the properties of a normal distribution. One approach you can use to determine whether a variable follows a normal distribution is to compare the observed characteristics of the variable with what would be expected if the variable followed a normal distribution. To do so, you can

- Construct charts and observe their appearance. For small- or moderate-sized data sets, create a stem-and-leaf display or a boxplot. For large data sets, in addition, plot a histogram or polygon.
- Compute descriptive statistics and compare these statistics with the theoretical properties of the normal distribution. Compare the mean and median. Is the interquartile range approximately 1.33 times the standard deviation? Is the range approximately 6 times the standard deviation?
- Evaluate how the values are distributed. Determine whether approximately two-thirds of the values lie between the mean and ± 1 standard deviation. Determine whether approximately four-fifths of the values lie between the mean and ± 1.28 standard deviations. Determine whether approximately 19 out of every 20 values lie between the mean and ± 2 standard deviations.

For example, you can use these techniques to determine whether the one-year returns discussed in Chapters 2 and 3 (stored in [Retirement Funds](#)) follow a normal distribution.

Table 6.6 presents the descriptive statistics and the five-number summary for the one-year return percentage variable. Figure 6.18 presents the Excel and Minitab boxplots for the one-year return percentages.

TABLE 6.6

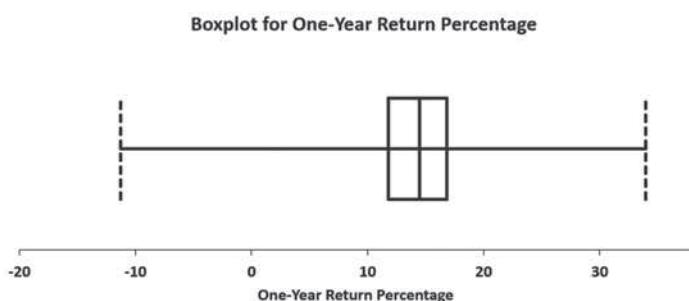
Descriptive Statistics and Five-Number Summary for the One-Year Return Percentages

Descriptive Statistics for 1YrReturn%			
Mean	14.40	Standard deviation	4.86
Median	14.48	Coeff. of variation	33.72%
Mode	14.50	Skewness	0.1036
Minimum	-11.28	Kurtosis	4.2511
Maximum	33.98	Count	316
Range	45.26	Standard error	0.27
Variance	23.57		

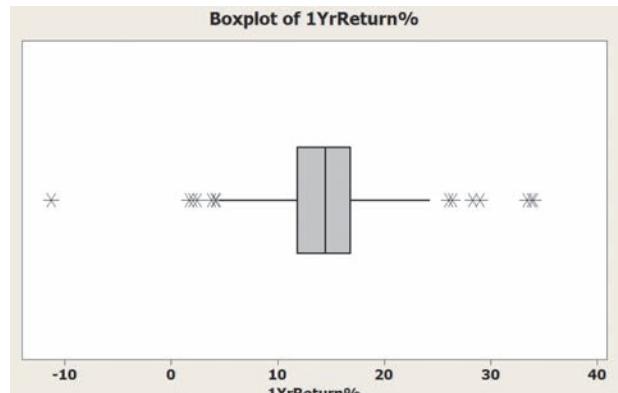
Five-Number Summary	
Minimum	-11.28
First quartile	11.80
Median	14.48
Third quartile	16.81
Maximum	33.98

FIGURE 6.18

Excel and Minitab boxplots for the one-year return percentages



Each of the two lines, or whiskers, that extend from the Minitab central box extend 1.5 times the interquartile range from the box. Beyond these ranges are the values considered to be outliers, plotted as asterisks.



From Table 6.6, Figure 6.18, and from an ordered array of the returns (not shown here), you can make the following statements about the one-year returns:

- The mean of 14.40 is approximately the same as the median of 14.48. (In a normal distribution, the mean and median are equal.)
- The boxplot is slightly left-skewed. (The normal distribution is symmetrical.)
- The interquartile range of 5.01 is approximately 1.03 standard deviations. (In a normal distribution, the interquartile range is 1.33 standard deviations.)
- The range of 45.26 is equal to 9.32 standard deviations. (In a normal distribution, the range is approximately 6 standard deviations.)
- 76.27% of the returns are within ± 1 standard deviation of the mean. (In a normal distribution, 68.26% of the values lie within ± 1 standard deviation of the mean.)
- 85.13% of the returns are within ± 1.28 standard deviations of the mean. (In a normal distribution, 80% of the values lie within ± 1.28 standard deviations of the mean.)
- 94.94% of the returns are within ± 2 standard deviations of the mean. (In a normal distribution, 95.44% of the values lie within ± 2 standard deviations of the mean.)
- The skewness statistic is 0.1036 and the kurtosis statistic is 4.2511. (In a normal distribution, each of these statistics equals zero.)

Based on these statements and the criteria given on page 233, you can conclude that the one-year returns are slightly skewed and have somewhat more values within ± 1 standard deviation of the mean than expected. The range is higher than what would be expected in a normal distribution, but this is mostly due to the single outlier at -11.28 . The skewness is very slightly positive, and the kurtosis indicates a distribution that is much more peaked than a normal distribution. Thus, you can conclude that the data characteristics of the one-year returns differ somewhat from the theoretical properties of a normal distribution.

Constructing the Normal Probability Plot

A **normal probability plot** is a visual display that helps you evaluate whether the data are normally distributed. One common plot is called the **quantile–quantile plot**. To create this plot, you first transform each ordered value to a Z value. For example, if you have a sample of $n = 19$, the Z value for the smallest value corresponds to a cumulative area of

$$\frac{1}{n + 1} = \frac{1}{19 + 1} = \frac{1}{20} = 0.05$$

The Z value for a cumulative area of 0.05 (from Table E.2) is -1.65 . Table 6.7 illustrates the entire set of Z values for a sample of $n = 19$.

TABLE 6.7

Ordered Values and Corresponding Z Values for a Sample of $n = 19$

Ordered Value	Z Value	Ordered Value	Z Value	Ordered Value	Z Value
1	-1.65	8	-0.25	14	0.52
2	-1.28	9	-0.13	15	0.67
3	-1.04	10	-0.00	16	0.84
4	-0.84	11	0.13	17	1.04
5	-0.67	12	0.25	18	1.28
6	-0.52	13	0.39	19	1.65
7	-0.39				

In a quantile–quantile plot, the Z values are plotted on the X axis, and the corresponding values of the variable are plotted on the Y axis. If the data are normally distributed, the values will plot along an approximately straight line. Figure 6.19 illustrates the typical shape of the quantile–quantile normal probability plot for a left-skewed distribution (Panel A), a normal distribution (Panel B), and a right-skewed distribution (Panel C). If the data are left-skewed, the curve will rise more rapidly at first and then level off. If the data are normally distributed, the points will plot along an approximately straight line. If the data are right-skewed, the data will rise more slowly at first and then rise at a faster rate for higher values of the variable being plotted.

FIGURE 6.19

Normal probability plots for a left-skewed distribution, a normal distribution, and a right-skewed distribution

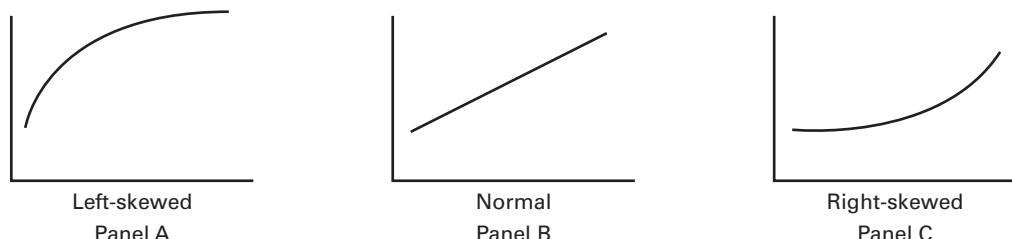
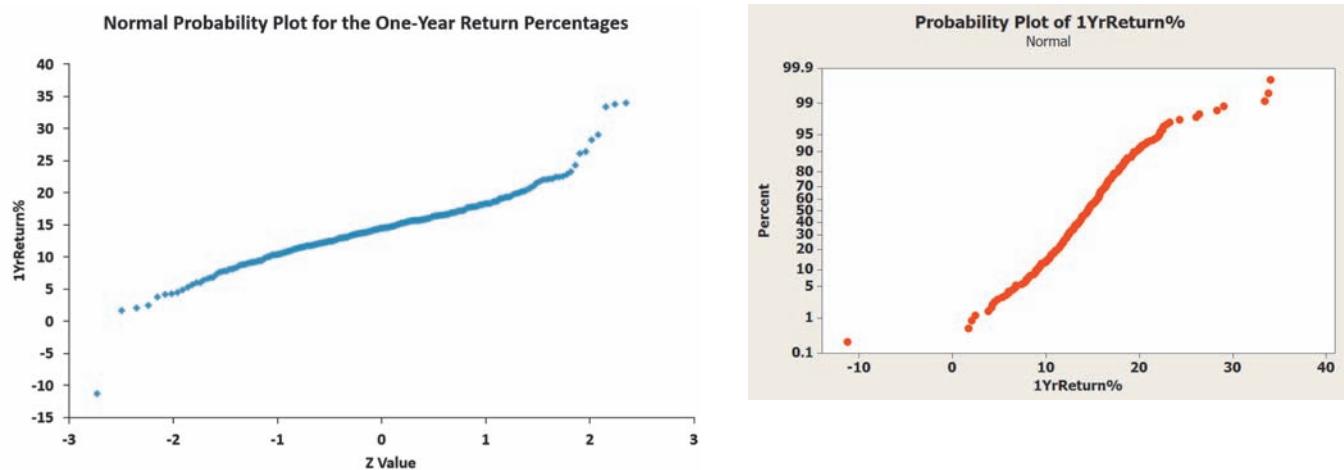


Figure 6.20 shows Excel (quantile–quantile) and Minitab normal probability plots for the one-year returns. The Excel quantile–quantile plot shows a single extremely low value followed by the bulk of the points that approximately follow a straight line except for a few high values.

FIGURE 6.20

Excel (quantile–quantile) and Minitab normal probability plots for the one-year returns



The Minitab normal probability plot has the one-year return percentage variable on the X axis and the cumulative percentage for a normal distribution on the Y axis. As with a quantile–quantile plot, the points will plot along an approximately straight line if the data are normally distributed. However, if the data are right-skewed, the curve will rise more rapidly at first and then level off. If the data are left-skewed, the data will rise more slowly at first and then rise at a faster rate for higher values of the variable being plotted. Observe that although the bulk of the points on the normal probability plot approximately follow a straight line, there are several high values that depart from a straight line, indicating a distribution that differs somewhat from a normal distribution.

Problems for Section 6.3

LEARNING THE BASICS

6.14 Show that for a sample of $n = 39$, the smallest and largest Z values are -1.96 and $+1.96$, and the middle (i.e., 20th) Z value is 0.00.

6.15 For a sample of $n = 6$, list the six Z values.

APPLYING THE CONCEPTS

6.16 The file **SUV** contains the overall miles per gallon (MPG) of 2013 small SUVs ($n = 17$):

22 23 21 22 25 26 22 22 21
19 22 22 26 23 24 21 22

Source: Data extracted from “Ratings,” *Consumer Reports*, April 2013, pp. 34–35.

Decide whether the data appear to be approximately normally distributed by

- comparing data characteristics to theoretical properties.
- constructing a normal probability plot.

6.17 As player salaries have increased, the cost of attending baseball games has increased dramatically. The file **BBCost2012**

contains the cost of four tickets, two beers, four soft drinks, four hot dogs, two game programs, two baseball caps, and the parking fee for one car for each of the 30 Major League Baseball teams in 2012:

176, 337, 223, 174, 233, 185, 160, 225, 324, 187, 196, 153, 184, 217, 146, 172, 300, 166, 184, 224, 213, 242, 172, 230, 257, 152, 225, 151, 224, 198

Source: Data extracted from fancostexperience.com/pages/fcx/fci_pdfs/8.pdf.

Decide whether the data appear to be approximately normally distributed by

- comparing data characteristics to theoretical properties.
- constructing a normal probability plot.

6.18 The file **Property Taxes** contains the property taxes per capita for the 50 states and the District of Columbia. Decide whether the data appear to be approximately normally distributed by

- comparing data characteristics to theoretical properties.
- constructing a normal probability plot.

6.19 Thirty companies comprise the DJIA. How big are these companies? One common method for measuring the size of a company is to use its market capitalization, which is computed by multiplying the number of stock shares by the price of a share of stock. On March 30, 2013, the market capitalization of these companies ranged from Alcoa's \$9.1 billion to ExxonMobil's \$403.7 billion. The entire population of market capitalization values is stored in **DowMarketCap**. (Data extracted from **money.cnn.com**, March 30, 2013.) Decide whether the market capitalization of companies in the DJIA appears to be approximately normally distributed by

- comparing data characteristics to theoretical properties.
- constructing a normal probability plot.
- constructing a histogram.

6.20 One operation of a mill is to cut pieces of steel into parts that will later be used as the frame for front seats in an automotive plant. The steel is cut with a diamond saw, and the resulting parts must be within ± 0.005 inch of the length specified by the automobile company. The data come from a sample of 100 steel parts and are stored in **Steel**. The measurement reported is the difference, in inches, between the actual length of the steel part, as measured by a laser measurement device, and the specified length of the steel part. Determine whether the data appear to be approximately normally distributed by

- comparing data characteristics to theoretical properties.
- constructing a normal probability plot.

6.21 The file **CD Rate** contains the yields for a one-year certificate of deposit (CD) and a five-year CD for 23 banks in the United States, as of March 20, 2013. (Data extracted from **www.Bankrate.com**, March 20, 2013.) For each type of investment, decide whether the data appear to be approximately normally distributed by

- comparing data characteristics to theoretical properties.
- constructing a normal probability plot.

6.22 The file **Utility** contains the electricity costs, in dollars, during July 2013 for a random sample of 50 one-bedroom apartments in a large city:

96	171	202	178	147	102	153	197	127	82
157	185	90	116	172	111	148	213	130	165
141	149	206	175	123	128	144	168	109	167
95	163	150	154	130	143	187	166	139	149
108	119	183	151	114	135	191	137	129	158

Decide whether the data appear to be approximately normally distributed by

- comparing data characteristics to theoretical properties.
- constructing a normal probability plot.

6.4 The Uniform Distribution

In the **uniform distribution**, the values are evenly distributed in the range between the smallest value, a , and the largest value, b . Because of its shape, the uniform distribution is sometimes called the **rectangular distribution** (see Panel B of Figure 6.1 on page 220). Equation (6.4) defines the probability density function for the uniform distribution.

UNIFORM PROBABILITY DENSITY FUNCTION

$$f(X) = \frac{1}{b-a} \text{ if } a \leq X \leq b \text{ and } 0 \text{ elsewhere} \quad (6.4)$$

where

$$\begin{aligned} a &= \text{minimum value of } X \\ b &= \text{maximum value of } X \end{aligned}$$

Equation (6.5) defines the mean of the uniform distribution, and Equation (6.6) defines the variance and standard deviation of the uniform distribution.

MEAN OF THE UNIFORM DISTRIBUTION

$$\mu = \frac{a+b}{2} \quad (6.5)$$

VARIANCE AND STANDARD DEVIATION OF THE UNIFORM DISTRIBUTION

$$\sigma^2 = \frac{(b - a)^2}{12} \quad (6.6a)$$

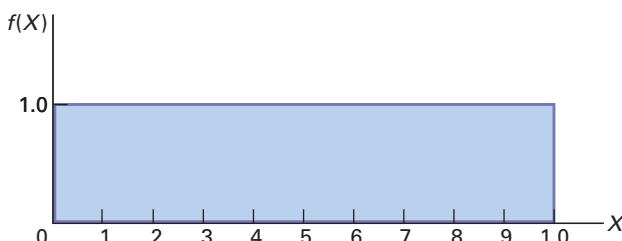
$$\sigma = \sqrt{\frac{(b - a)^2}{12}} \quad (6.6b)$$

One of the most common uses of the uniform distribution is in the selection of random numbers. When you use simple random sampling (see Section 1.4), you assume that each random digit comes from a uniform distribution that has a minimum value of 0 and a maximum value of 9.

Figure 6.21 illustrates the uniform distribution with $a = 0$ and $b = 1$. The total area inside the rectangle is equal to the base (1.0) times the height (1.0). Thus, the resulting area of 1.0 satisfies the requirement that the area under any probability density function equals 1.0.

FIGURE 6.21

Probability density function for a uniform distribution with $a = 0$ and $b = 1$

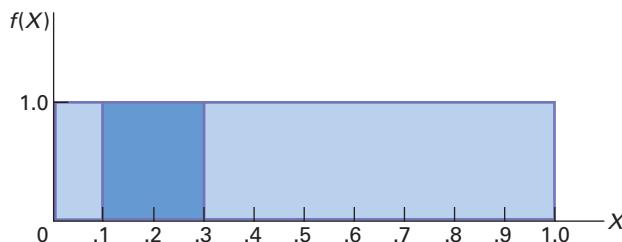


In this uniform distribution, what is the probability of getting a random number between 0.10 and 0.30? The area between 0.10 and 0.30, depicted in Figure 6.22, is equal to the base (which is $0.30 - 0.10 = 0.20$) times the height (1.0). Therefore,

$$P(0.10 < X < 0.30) = (\text{Base})(\text{Height}) = (0.20)(1.0) = 0.20$$

FIGURE 6.22

Finding $P(0.10 < X < 0.30)$ for a uniform distribution with $a = 0$ and $b = 1$



From Equations (6.5) and (6.6), the mean and standard deviation of the uniform distribution for $a = 0$ and $b = 1$ are computed as follows:

$$\begin{aligned}\mu &= \frac{a + b}{2} \\ &= \frac{0 + 1}{2} = 0.5\end{aligned}$$

and

$$\begin{aligned}\sigma^2 &= \frac{(b-a)^2}{12} \\ &= \frac{(1-0)^2}{12} \\ &= \frac{1}{12} = 0.0833 \\ \sigma &= \sqrt{0.0833} = 0.2887\end{aligned}$$

Thus, the mean is 0.5, and the standard deviation is 0.2887.

Example 6.6 provides another application of the uniform distribution.

EXAMPLE 6.6

Computing Uniform Probabilities

In the Normal Downloading at MyTVLab scenario on page 219, the download time of videos was assumed to be normally distributed with a mean of 7 seconds. Suppose that the download time follows a uniform (instead of a normal) distribution between 4.5 and 9.5 seconds. What is the probability that a download time will take more than 9 seconds?

SOLUTION The download time is uniformly distributed from 4.5 to 9.5 seconds. The area between 9 and 9.5 seconds is equal to 0.5 seconds, and the total area in the distribution is $9.5 - 4.5 = 5$ seconds. Therefore, the probability of a download time between 9 and 9.5 seconds is the portion of the area greater than 9, which is equal to $0.5/5.0 = 0.10$. Because 9.5 is the maximum value in this distribution, the probability of a download time above 9 seconds is 0.10. In comparison, if the download time is normally distributed with a mean of 7 seconds and a standard deviation of 2 seconds (see Example 6.1 on page 225), the probability of a download time above 9 seconds is 0.1587.

Problems for Section 6.4

LEARNING THE BASICS

6.23 Suppose you select one value from a uniform distribution with $a = 0$ and $b = 10$. What is the probability that the value will be

- a. between 5 and 7?
- b. between 2 and 3?
- c. What is the mean?
- d. What is the standard deviation?

APPLYING THE CONCEPTS

SELF Test **6.24** The time between arrivals of customers at a bank during the noon-to-1 P.M. hour has a uniform distribution between 0 to 120 seconds. What is the probability that the time between the arrival of two customers will be

- a. less than 20 seconds?
- b. between 10 and 30 seconds?
- c. more than 35 seconds?
- d. What are the mean and standard deviation of the time between arrivals?

6.25 A study of the time spent shopping in a supermarket for a market basket of 20 specific items showed an approximately uniform distribution between 20 minutes and 40 minutes. What is the probability that the shopping time will be

- a. between 25 and 30 minutes?
- b. less than 35 minutes?
- c. What are the mean and standard deviation of the shopping time?

6.26 How long does it take to download a two-hour movie from the iTunes store? According to Apple's technical support site, support.apple.com/kb/ht1577, downloading such a movie using a 5 Mbit/s broadband connection should take 18 to 24 minutes. Assume that the download times are uniformly distributed between 18 and 24 minutes. If you download a two-hour movie, what is the probability that the download time will be

- a. less than 19 minutes?
- b. more than 23 minutes?
- c. between 20 and 22 minutes?
- d. What are the mean and standard deviation of the download times?

6.27 The scheduled commuting time on the Long Island Railroad from Glen Cove to New York City is 65 minutes. Suppose that the actual commuting time is uniformly distributed between 64 and 74 minutes. What is the probability that the commuting time will be

- a. less than 70 minutes?
- b. between 65 and 70 minutes?
- c. greater than 65 minutes?
- d. What are the mean and standard deviation of the commuting time?

6.5 The Exponential Distribution

The **exponential distribution** is a continuous distribution that is right-skewed and ranges from zero to positive infinity (see Panel C of Figure 6.1 on page 220). The exponential distribution is widely used in waiting-line (i.e., queuing) theory to model the length of time between arrivals in processes such as customers arriving at a bank's ATM, patients entering a hospital emergency room, and hits on a website.

The exponential distribution is defined by a single parameter, λ , the mean number of arrivals per unit of time. The probability density function for the length of time between arrivals is given by Equation (6.7).

EXPONENTIAL PROBABILITY DENSITY FUNCTION

$$f(X) = \lambda e^{-\lambda x} \text{ for } X > 0 \quad (6.7)$$

where

e = mathematical constant approximated by 2.71828

λ = mean number of arrivals per unit

X = any value of the continuous variable where $0 < X < \infty$

The mean time between arrivals, μ , is given by Equation (6.8), and the standard deviation of the time between arrivals, σ , is given by Equation (6.9).

MEAN TIME BETWEEN ARRIVALS

$$\mu = \frac{1}{\lambda} \quad (6.8)$$

STANDARD DEVIATION OF THE TIME BETWEEN ARRIVALS

$$\sigma = \frac{1}{\lambda} \quad (6.9)$$

The value $1/\lambda$ is equal to the mean time between arrivals. For example, if the mean number of arrivals in a minute is $\lambda = 4$, then the mean time between arrivals is $1/\lambda = 0.25$ minute, or 15 seconds. Equation (6.10) defines the cumulative probability that the length of time before the next arrival is less than or equal to X .

CUMULATIVE EXPONENTIAL PROBABILITY

$$P(\text{arrival time} \leq X) = 1 - e^{-\lambda x} \quad (6.10)$$

To illustrate the exponential distribution, suppose that customers arrive at a bank's ATM at a rate of 20 per hour. If a customer has just arrived, what is the probability that the next customer will arrive within 6 minutes (i.e., 0.1 hour)? For this example, $\lambda = 20$ and $X = 0.1$. Using Equation (6.10),

$$\begin{aligned} P(\text{arrival time} \leq 0.1) &= 1 - e^{-20(0.1)} \\ &= 1 - e^{-2} \\ &= 1 - 0.1353 = 0.8647 \end{aligned}$$

Thus, the probability that a customer will arrive within 6 minutes is 0.8647, or 86.47%. Figure 6.23 shows this probability as computed by Excel and Minitab.

FIGURE 6.23

Excel and Minitab results for computing exponential probability that a customer will arrive within six minutes

A	B
1	Exponential Probability
2	
3	Data
4	Mean 20
5	X Value 0.1
6	
7	Results
8	P($\leq X$) 0.8647 =EXPO.DIST(B5, B4, TRUE)

Cumulative Distribution Function
Exponential with mean = 0.05
 $x \ P(X \leq x)$
0.1 0.864665

Example 6.7 illustrates the effect on the exponential probability of changing the time between arrivals.

EXAMPLE 6.7

Computing Exponential Probabilities

In the ATM example, what is the probability that the next customer will arrive within 3 minutes (i.e., 0.05 hour)?

SOLUTION For this example, $\lambda = 20$ and $X = 0.05$. Using Equation (6.10),

$$\begin{aligned} P(\text{arrival time} \leq 0.05) &= 1 - e^{-20(0.05)} \\ &= 1 - e^{-1} \\ &= 1 - 0.3679 = 0.6321 \end{aligned}$$

Thus, the probability that a customer will arrive within 3 minutes is 0.6321, or 63.21%.

Problems for Section 6.5

LEARNING THE BASICS

6.28 Given an exponential distribution with $\lambda = 10$, what is the probability that the arrival time is

- a. less than $X = 0.1$?
- b. greater than $X = 0.1$?
- c. between $X = 0.1$ and $X = 0.2$?
- d. less than $X = 0.1$ or greater than $X = 0.2$?

6.29 Given an exponential distribution with $\lambda = 30$, what is the probability that the arrival time is

- a. less than $X = 0.1$?
- b. greater than $X = 0.1$?
- c. between $X = 0.1$ and $X = 0.2$?
- d. less than $X = 0.1$ or greater than $X = 0.2$?

6.30 Given an exponential distribution with $\lambda = 5$, what is the probability that the arrival time is

- a. less than $X = 0.3$?
- b. greater than $X = 0.3$?
- c. between $X = 0.3$ and $X = 0.5$?
- d. less than $X = 0.3$ or greater than $X = 0.5$?

APPLYING THE CONCEPTS

6.31 Autos arrive at a toll plaza located at the entrance to a bridge at a rate of 50 per minute during the 5:00-to-6:00 p.m. hour. If an auto has just arrived,

- a. what is the probability that the next auto will arrive within 3 seconds (0.05 minute)?
- b. what is the probability that the next auto will arrive within 1 second (0.0167 minute)?
- c. What are your answers to (a) and (b) if the rate of arrival of autos is 60 per minute?
- d. What are your answers to (a) and (b) if the rate of arrival of autos is 30 per minute?

6.32 Customers arrive at the drive-up window of a fast-food restaurant at a rate of 2 per minute during the lunch hour.

- a. What is the probability that the next customer will arrive within 1 minute?
- b. What is the probability that the next customer will arrive within 5 minutes?
- c. During the dinner time period, the arrival rate is 1 per minute. What are your answers to (a) and (b) for this period?

6.33 Telephone calls arrive at the information desk of a large computer software company at a rate of 15 per hour.

- What is the probability that the next call will arrive within 3 minutes (0.05 hour)?
- What is the probability that the next call will arrive within 15 minutes (0.25 hour)?
- Suppose the company has just introduced an updated version of one of its software programs, and telephone calls are now arriving at a rate of 25 per hour. Given this information, what are your answers to (a) and (b)?

6.34 Calls arrive at a call center at the rate of 12 per hour. What is the probability that the next call arrives in

- less than 3 minutes?
- more than 6 minutes?
- less than 1 minute?

6.35 The time between unplanned shutdowns of a power plant has an exponential distribution with a mean of 20 days. Find the probability that the time between two unplanned shutdowns is

- less than 14 days.
- more than 21 days.
- less than 7 days.

6.36 Golfers arrive at the starter's booth of a public golf course at a rate of 8 per hour during the Monday-to-Friday midweek period. If a golfer has just arrived,

- what is the probability that the next golfer will arrive within 15 minutes (0.25 hour)?
- what is the probability that the next golfer will arrive within 3 minutes (0.05 hour)?
- The actual arrival rate on Fridays is 15 per hour. What are your answers to (a) and (b) for Fridays?

6.37 Some Internet companies sell a service that will boost a website's traffic by delivering additional unique visitors. Assume that one such company claims it can deliver 1,000 visitors a day. If this amount of website traffic is experienced, then the time between visitors has a mean of 1.44 minutes (or 0.6944 per minute). Assume that your website gets 1,000 visitors a day and that the time between visitors has an exponential distribution. What is the probability that the time between two visitors is

- less than 1 minute?
- less than 2 minutes?
- more than 3 minutes?
- Do you think it is reasonable to assume that the time between visitors has an exponential distribution?

6.6 The Normal Approximation to the Binomial Distribution

In many circumstances, you can use the normal distribution to approximate the binomial distribution that is discussed in Section 5.3. The **Section 6.6 online topic** explains this approximation and illustrates its use.

USING STATISTICS

Normal Downloading at MyTVLab, Revisited

In the Normal Downloading at MyTVLab scenario, you were a project manager for an online social media and video website. You sought to ensure that a video could be downloaded quickly by visitors to the website. By running experiments in the corporate offices, you determined that the amount of time, in seconds, that passes from clicking a download link until a video is fully displayed is a bell-shaped distribution with a mean download time of 7 seconds and standard deviation of 2 seconds. Using the normal distribution, you were able to calculate that approximately 84% of the download times are 9 seconds or less, and 95% of the download times are between 3.08 and 10.92 seconds.

Now that you understand how to compute probabilities from the normal distribution, you can evaluate download times of a video using different website designs. For example, if the standard deviation remained at 2 seconds, lowering the mean to 6 seconds would shift the entire distribution lower by



Cloki/Shutterstock

1 second. Thus, approximately 84% of the download times would be 8 seconds or less, and 95% of the download times would be between 2.08 and 9.92 seconds. Another change that could reduce long download times would be reducing the variation. For example, consider the case where the mean remained at the original 7 seconds but the standard deviation was reduced to 1 second. Again, approximately 84% of the download times would be 8 seconds or less, and 95% of the download times would be between 5.04 and 8.96 seconds.

SUMMARY

In this and the previous chapter, you have learned about mathematical models called probability distributions and how they can be used to solve business problems. In Chapter 5, you used discrete probability distributions in situations where the values come from a counting process such as the number of social media sites to which you belong or the number of tagged order forms in a report generated by an accounting information system. In this chapter, you learned about continuous probability distributions where the values come from a measuring process such as your height or the download time of a video.

Continuous probability distributions come in various shapes, but the most common and most important in business is the normal distribution. The normal distribution is symmetrical; thus, its mean and median are equal. It is

also bell-shaped, and approximately 68.26% of its values are within ± 1 standard deviation of the mean, approximately 95.44% of its values are within ± 2 standard deviations of the mean, and approximately 99.73% of its values are within ± 3 standard deviations of the mean. Although many variables in business are closely approximated by the normal distribution, do not think that all variables can be approximated by the normal distribution.

In Section 6.3, you learned about various methods for evaluating normality in order to determine whether the normal distribution is a reasonable mathematical model to use in specific situations. In Sections 6.4 and 6.5, you studied other continuous distributions—in particular, the uniform and exponential distributions. Chapter 7 uses the normal distribution to develop the subject of statistical inference.

REFERENCES

1. Gunter, B. "Q-Q Plots." *Quality Progress* (February 1994): 81–86.
2. Levine, D. M., P. Ramsey, and R. Smidt. *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.
3. *Microsoft Excel 2013*. Redmond, WA: Microsoft Corp., 2012.
4. Miller, J. "Earliest Known Uses of Some of the Words of Mathematics." jeff560.tripod.com/mathword.html.
5. *Minitab Release 16*. State College, PA: Minitab, Inc., 2010.
6. Pearl, R. "Karl Pearson, 1857–1936." *Journal of the American Statistical Association*, 31 (1936): 653–664.
7. Pearson, E. S. "Some Incidents in the Early History of Biometry and Statistics, 1890–94." *Biometrika* 52 (1965): 3–18.
8. Taleb, N. *The Black Swan*, 2nd ed. New York: Random House, 2010.
9. Walker, H. "The Contributions of Karl Pearson." *Journal of the American Statistical Association* 53 (1958): 11–22.

KEY EQUATIONS

Normal Probability Density Function

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)[(X-\mu)/\sigma]^2} \quad (6.1)$$

Z Transformation Formula

$$Z = \frac{X - \mu}{\sigma} \quad (6.2)$$

Finding an X Value Associated with a Known Probability

$$X = \mu + Z\sigma \quad (6.3)$$

Uniform Probability Density Function

$$f(X) = \frac{1}{b-a} \quad (6.4)$$

Mean of the Uniform Distribution

$$\mu = \frac{a+b}{2} \quad (6.5)$$

Variance and Standard Deviation of the Uniform Distribution

$$\sigma^2 = \frac{(b-a)^2}{12} \quad (6.6a)$$

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} \quad (6.6b)$$

Exponential Probability Density Function

$$f(X) = \lambda e^{-\lambda x} \text{ for } X > 0 \quad (6.7)$$

Mean Time Between Arrivals

$$\mu = \frac{1}{\lambda} \quad (6.8)$$

Standard Deviation of the Time Between Arrivals

$$\sigma = \frac{1}{\lambda} \quad (6.9)$$

Cumulative Exponential Probability

$$P(\text{arrival time} \leq X) = 1 - e^{-\lambda x} \quad (6.10)$$

KEY TERMS

cumulative standardized normal distribution 223
 exponential distribution 240
 normal distribution 220
 normal probability plot 235

probability density function 220
 probability density function for the normal distribution 222
 quantile-quantile plot 235
 rectangular distribution 237

standardized normal variable 222
 transformation formula 222
 uniform distribution 237

CHECKING YOUR UNDERSTANDING

6.38 Why is only one normal distribution table such as Table E.2 needed to find any probability under the normal curve?

6.39 How do you find the area between two values under the normal curve?

6.40 How do you find the X value that corresponds to a given percentile of the normal distribution?

6.41 What are some of the distinguishing properties of a normal distribution?

6.42 How does the shape of the normal distribution differ from the shapes of the uniform and exponential distributions?

6.43 How can you use the normal probability plot to evaluate whether a set of data is normally distributed?

6.44 Under what circumstances can you use the exponential distribution?

CHAPTER REVIEW PROBLEMS

6.45 An industrial sewing machine uses ball bearings that are targeted to have a diameter of 0.75 inch. The lower and upper specification limits under which the ball bearings can operate are 0.74 inch and 0.76 inch, respectively. Past experience has indicated that the actual diameter of the ball bearings is approximately normally distributed, with a mean of 0.753 inch and a standard deviation of 0.004 inch. What is the probability that a ball bearing is

- between the target and the actual mean?
- between the lower specification limit and the target?
- above the upper specification limit?
- below the lower specification limit?
- Of all the ball bearings, 93% of the diameters are greater than what value?

6.46 The fill amount in 2-liter soft drink bottles is normally distributed, with a mean of 2.0 liters and a standard deviation of 0.05 liter. If bottles contain less than 95% of the listed net content (1.90 liters, in this case), the manufacturer may be subject to penalty by the state office of consumer affairs. Bottles that have a net content above 2.10 liters may cause excess spillage upon opening. What proportion of the bottles will contain

- between 1.90 and 2.0 liters?
- between 1.90 and 2.10 liters?
- below 1.90 liters or above 2.10 liters?
- At least how much soft drink is contained in 99% of the bottles?
- Ninety-nine percent of the bottles contain an amount that is between which two values (symmetrically distributed) around the mean?

6.47 In an effort to reduce the number of bottles that contain less than 1.90 liters, the bottler in Problem 6.46 sets the filling machine so that the mean is 2.02 liters. Under these circumstances, what are your answers in Problem 6.46 (a) through (e)?

6.48 An Ipsos MediaCT study indicates that mobile device owners who used their mobile device while shopping for consumer

electronics spent an average of \$1,539 on consumer electronics in the past six months. (Data extracted from [iab.net/showrooming](#).) Assume that the amount spent on consumer electronics in the last six months is normally distributed and that the standard deviation is \$500.

- What is the probability that a mobile device owner who used his or her mobile device while shopping for consumer electronics spent less than \$1,000 on consumer electronics?
- What is the probability that a mobile device owner who used his or her mobile device while shopping for consumer electronics spent between \$2,500 and \$3,000 on consumer electronics?
- Ninety percent of the amounts spent on consumer electronics by mobile device owners who used their mobile device while shopping for consumer electronics are less than what value?
- Eighty percent of the amounts spent on consumer electronics by mobile device owners who used their mobile device while shopping for consumer electronics are between what two values symmetrically distributed around the mean?

6.49 The file [DomesticBeer](#) contains the percentage alcohol, number of calories per 12 ounces, and number of carbohydrates (in grams) per 12 ounces for 152 of the best-selling domestic beers in the United States. Determine whether each of these variables appears to be approximately normally distributed. Support your decision through the use of appropriate statistics and graphs. (Data extracted from [www.Beer100.com](#), March 20, 2013.)

6.50 The evening manager of a restaurant was very concerned about the length of time some customers were waiting in line to be seated. She also had some concern about the seating times—that is, the length of time between when a customer is seated and the time he or she leaves the restaurant. Over the course of one week, 100 customers (no more than 1 per party) were randomly selected, and their waiting and seating times (in minutes) were recorded in [Wait](#).

- a. Think about your favorite restaurant. Do you think waiting times more closely resemble a uniform, an exponential, or a normal distribution?
- b. Again, think about your favorite restaurant. Do you think seating times more closely resemble a uniform, an exponential, or a normal distribution?
- c. Construct a histogram and a normal probability plot of the waiting times. Do you think these waiting times more closely resemble a uniform, an exponential, or a normal distribution?
- d. Construct a histogram and a normal probability plot of the seating times. Do you think these seating times more closely resemble a uniform, an exponential, or a normal distribution?

6.51 The major stock market indexes had strong results in 2012. The mean one-year return for stocks in the S&P 500, a group of 500 very large companies, was +13.41%. The mean one-year return for the NASDAQ, a group of 3,200 small and medium-sized companies, was +15.91%. Historically, the one-year returns are approximately normally distributed, the standard deviation in the S&P 500 is approximately 20%, and the standard deviation in the NASDAQ is approximately 30%.

- a. What is the probability that a stock in the S&P 500 gained value in 2012?
- b. What is the probability that a stock in the S&P 500 gained 10% or more in 2012?
- c. What is the probability that a stock in the S&P 500 lost 20% or more in 2012?
- d. What is the probability that a stock in the S&P 500 lost 30% or more in 2012?
- e. Repeat (a) through (d) for a stock in the NASDAQ.
- f. Write a short summary on your findings. Be sure to include a discussion of the risks associated with a large standard deviation.

6.52 The speed at which you can log into a website through a smartphone is an important quality characteristic of that website. In a recent test, the mean time to log into the JetBlue Airways website through a smartphone was 4.237 seconds. (Data extracted from N. Trejos, "Travelers Have No Patience for Slow Mobile Sites," *USA Today*, April 4, 2012, p. 3B.) Suppose that the download time is normally distributed, with a standard deviation of 1.3 seconds. What is the probability that a download time is

- a. less than 2 seconds?
- b. between 1.5 and 2.5 seconds?
- c. above 1.8 seconds?
- d. Ninety-nine percent of the download times are slower (higher) than how many seconds?
- e. Ninety-five percent of the download times are between what two values, symmetrically distributed around the mean?

- f. Suppose that the download times are uniformly distributed between 1 and 9 seconds. What are your answers to (a) through (c)?

6.53 The speed at which you can log into a website through a smartphone is an important quality characteristic of that website. In a recent test, the mean time to log into the Hertz website through a smartphone was 7.524 seconds. (Data extracted from N. Trejos, "Travelers Have No Patience for Slow Mobile Sites," *USA Today*, April 4, 2012, p. 3B.) Suppose that the download time is normally distributed, with a standard deviation of 1.7 seconds. What is the probability that a download time is

- a. less than 2 seconds?
- b. between 1.5 and 2.5 seconds?
- c. above 1.8 seconds?
- d. Ninety-nine percent of the download times are slower (higher) than how many seconds?
- e. Ninety-five percent of the download times are between what two values, symmetrically distributed around the mean?
- f. Suppose that the download times are uniformly distributed between 1 and 14 seconds. What are your answers to (a) through (d)?
- g. Compare the results for the JetBlue Airways site computed in Problem 6.52 to those of the Hertz website.

6.54 (Class Project) One theory about the daily changes in the closing price of stock is that these changes follow a *random walk*—that is, these daily events are independent of each other and move upward or downward in a random manner—and can be approximated by a normal distribution. To test this theory, use either a newspaper or the Internet to select one company traded on the NYSE, one company traded on the American Stock Exchange, and one company traded on the NASDAQ and then do the following:

1. Record the daily closing stock price of each of these companies for six consecutive weeks (so that you have 30 values per company).
2. Compute the daily changes in the closing stock price of each of these companies for six consecutive weeks (so that you have 30 values per company).

Note: The random-walk theory pertains to the daily changes in the closing stock price, not the daily closing stock price.

For each of your six data sets, decide whether the data are approximately normally distributed by

- a. constructing the stem-and-leaf display, histogram or polygon, and boxplot.
- b. comparing data characteristics to theoretical properties.
- c. constructing a normal probability plot.
- d. Discuss the results of (a) through (c). What can you say about your three stocks with respect to daily closing prices and daily changes in closing prices? Which, if any, of the data sets are approximately normally distributed?

CASES FOR CHAPTER 6

Managing Ashland MultiComm Services

The AMS technical services department has embarked on a quality improvement effort. Its first project relates to maintaining the target upload speed for its Internet service subscribers. Upload speeds are measured on a standard scale in which the target value is 1.0. Data collected over the past

year indicate that the upload speed is approximately normally distributed, with a mean of 1.005 and a standard deviation of 0.10. Each day, one upload speed is measured. The upload speed is considered acceptable if the measurement on the standard scale is between 0.95 and 1.05.

1. Assuming that the distribution has not changed from what it was in the past year, what is the probability that the upload speed is
 - a. less than 1.0?
 - b. between 0.95 and 1.0?
 - c. between 1.0 and 1.05?
 - d. less than 0.95 or greater than 1.05?
2. The objective of the operations team is to reduce the probability that the upload speed is below 1.0. Should the team focus on process improvement that increases the mean upload speed to 1.05 or on process improvement that reduces the standard deviation of the upload speed to 0.075? Explain.

Digital Case

Apply your knowledge about the normal distribution in this Digital Case, which extends the Using Statistics scenario from this chapter.

To satisfy concerns of potential customers, the management of MyTVLab has undertaken a research project to learn how much time it takes users to load a complex video features page. The research team has collected data and has made some claims based on the assertion that the data follow a normal distribution.

Open [MTL_QRTStudy.pdf](#), which documents the work of a quality response team at MyTVLab. Read the

internal report that documents the work of the team and their conclusions. Then answer the following:

1. Can the collected data be approximated by the normal distribution?
2. Review and evaluate the conclusions made by the MyTVLab research team. Which conclusions are correct? Which ones are incorrect?
3. If MyTVLab could improve the mean time by five seconds, how would the probabilities change?

CardioGood Fitness

Return to the CardioGood Fitness case (stored in [CardioGood Fitness](#)) first presented on page 83.

1. For each CardioGood Fitness treadmill product line, determine whether the age, income, usage, and the

number of miles the customer expects to walk/run each week can be approximated by the normal distribution.

2. Write a report to be presented to the management of CardioGood Fitness, detailing your findings.

More Descriptive Choices Follow-up

Follow up the More Descriptive Choices Revisited Using Statistics scenario on page 144 by constructing normal probability plots for the 3-year return percentages, 5-year return percentages, and 10-year return percentages for the

sample of 316 retirement funds stored in [Retirement Funds](#). In your analysis, examine differences between the growth and value funds as well as the differences among the small, mid-cap, and large market cap funds.

Clear Mountain State Student Surveys

1. The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students who attend CMSU. They create and distribute a survey of 14 questions and receive responses from 62 undergraduates (stored in [UndergradSurvey](#)). For each numerical variable in the survey, decide whether the variable is approximately normally distributed by
 - a. comparing data characteristics to theoretical properties.
 - b. constructing a normal probability plot.
 - c. writing a report summarizing your conclusions.
2. The dean of students at CMSU has learned about the undergraduate survey and has decided to undertake a similar survey for graduate students at CMSU. She creates and distributes a survey of 14 questions and receives responses from 44 graduate students (stored in [GradSurvey](#)). For each numerical variable in the survey, decide whether the variable is approximately normally distributed by
 - a. comparing data characteristics to theoretical properties.
 - b. constructing a normal probability plot.
 - c. writing a report summarizing your conclusions.

CHAPTER 6 EXCEL GUIDE

EG6.1 CONTINUOUS PROBABILITY DISTRIBUTIONS

There are no Excel Guide instructions for this section.

EG6.2 The NORMAL DISTRIBUTION

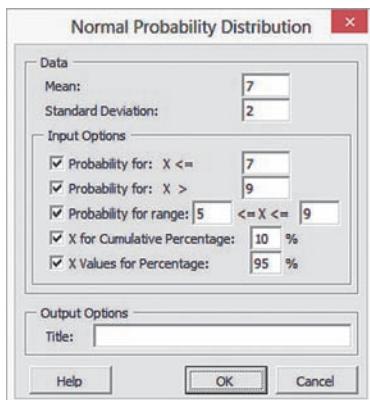
Key Technique Use the **NORM.DIST(X value, mean, standard deviation, True)** function to compute normal probabilities and use the **NORM.S.INV(percentage)** function and the STANDARDIZE function (see Section EG3.2) to compute the Z value.

Example Compute the normal probabilities for Examples 6.1 through 6.3 on pages 225 and 226 and the X and Z values for Examples 6.4 and 6.5 on pages 228 and 229.

PHStat Use Normal.

For the example, select **PHStat → Probability & Prob. Distributions → Normal**. In this procedure's dialog box (shown below):

1. Enter 7 as the **Mean** and 2 as the **Standard Deviation**.
2. Check **Probability for: X <=** and enter 7 in its box.
3. Check **Probability for: X >** and enter 9 in its box.
4. Check **Probability for range** and enter 5 in the first box and 9 in the second box.
5. Check **X for Cumulative Percentage** and enter 10 in its box.
6. Check **X Values for Percentage** and enter 95 in its box.
7. Enter a **Title** and click **OK**.



In-Depth Excel Use the **COMPUTE worksheet** of the **Normal workbook** as a template.

The worksheet already contains the data for solving the problems in Examples 6.1 through 6.5. For other problems, change the values for the **Mean**, **Standard Deviation**, **X Value**, **From X Value**, **To X Value**, **Cumulative Percentage**, and/or **Percentage**.

Read the **SHORT TAKES** for Chapter 6 for an explanation of the formulas found in the **COMPUTE worksheet** (shown in the **COMPUTE_FORMULAS worksheet**). If you use an Excel version older than Excel 2010, use the **COMPUTE_Older worksheet**.

EG6.3 EVALUATING NORMALITY

Comparing Data Characteristics to Theoretical Properties

Use the Sections EG3.1 through EG3.3 instructions to compare data characteristics to theoretical properties.

Constructing the Normal Probability Plot

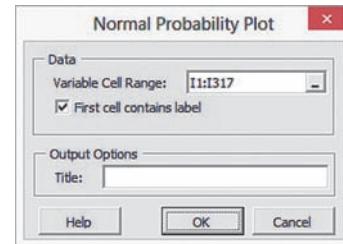
Key Technique Use an Excel Scatter (X, Y) chart with Z values computed using the **NORM.S.INV** function.

Example Construct the normal probability plot for the one-year return percentages for the sample of 316 retirement funds that is shown in Figure 6.20 on page 236.

PHStat Use Normal Probability Plot.

For the example, open to the **DATA worksheet** of the **Retirement Funds workbook**. Select **PHStat → Probability & Prob. Distributions → Normal Probability Plot**. In the procedure's dialog box (shown below):

1. Enter **I1:I317** as the **Variable Cell Range**.
2. Check **First cell contains label**.
3. Enter a **Title** and click **OK**.



In addition to the chart sheet containing the normal probability plot, the procedure creates a plot data worksheet identical to the **PlotData** worksheet discussed in the **In-Depth Excel** instructions.

In-Depth Excel Use the worksheets of the **NPP workbook** as templates.

The **NormalPlot** chart sheet displays a normal probability plot using the rank, the proportion, the Z value, and the variable found in the **PLOT_DATA worksheet**. The **PLOT_DATA** worksheet already contains the one-year return percentages for the example. To construct a plot for a different variable, paste the *sorted* values for that variable in **column D** of the **PLOT_DATA worksheet**. Adjust the number of ranks in **column A** and the divisor in the formulas in **column B** to compute cumulative percentages to reflect the quantity $n + 1$ (317 for the example). (Column C formulas use the **NORM.S.INV** function to compute the Z values for those cumulative percentages.)

If you have fewer than 316 values, delete rows from the bottom up. If you have more than 316 values, select row 317, right-click, click **Insert** in the shortcut menu, and copy down the formulas in columns B and C to the new rows. To create your own

normal probability plot for the 1YrReturn% variable, open to the PLOT_DATA worksheet and select the cell range C1:D317. Then select **Insert → Scatter** and select the first Scatter gallery item (that shows only points and is labeled with **Scatter** or **Scatter with only Markers**). Relocate the chart to a chart sheet, turn off the chart legend and gridlines, add axis titles, and modify the chart title by using the instructions in Appendix Section B.6.

If you use an Excel version older than Excel 2010, use the PLOT_Older worksheet and the NormalPlot_Older chart sheet.

EG6.4 The UNIFORM DISTRIBUTION

There are no Excel Guide instructions for this section.

EG6.5 The EXPONENTIAL DISTRIBUTION

Key Technique Use the EXPON.DIST(*X value, mean, True*) function.

Example Compute the exponential probability for the bank ATM customer arrival example on page 240.

CHAPTER 6 MINITAB GUIDE

MG6.1 CONTINUOUS PROBABILITY DISTRIBUTIONS

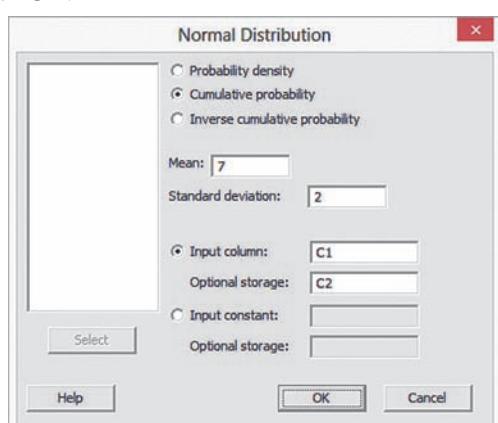
There are no Minitab Guide instructions for this section.

MG6.2 The NORMAL DISTRIBUTION

Use **Normal**.

For example, to compute the normal probability for Example 6.1 on page 225, open to a new worksheet. Enter **X Value** as the name of column **C1** and enter **9** in the row 1 cell of that column. Select **Calc → Probability Distributions → Normal**. In the Normal Distribution dialog box (shown below):

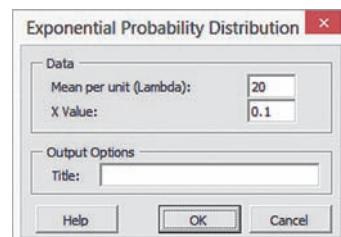
1. Click **Cumulative probability**.
2. Enter **7** in the **Mean** box.
3. Enter **2** in the **Standard deviation** box.
4. Click **Input column** and enter **C1** in its box and press **Tab**.
5. Enter **C2** in the first **Optional storage** box.
6. Click **OK**.



PHStat Use Exponential.

For the example, select **PHStat → Probability & Prob. Distributions → Exponential**. In the procedure's dialog box (shown below):

1. Enter **20** as the **Mean per unit (Lambda)** and **0.1** as the **X Value**.
2. Enter a **Title** and click **OK**.



In-Depth Excel Use the COMPUTE worksheet of the Exponential workbook as a template.

The worksheet already contains the data for the example. For other problems, change the **Mean** and **X Value** in cells **B4** and **B5**. If you use an Excel version older than Excel 2010, use the COMPUTE_Older worksheet.

Minitab places in the row 1 cell of column C2 the probability for a download time that is *less than* 9 seconds with $\mu = 7$ and $\sigma = 2$. To compute the Example 6.1 probability for a download time that is *greater than* 9 seconds, select **Calc → Calculator**. Enter C3 in the **Store result in variable** box, enter $1 - C2$ in the **Expression** box, and click **OK**. The probability appears in row 1 of column C3.

To compute the normal probability for Example 6.4 on page 228, open to a new worksheet. Enter **Cumulative Percentage** as the name of column **C1** and enter **0.1** in the row 1 cell of that column. Select **Calc → Probability Distributions → Normal**. In the Normal Distribution dialog box:

1. Click **Inverse cumulative probability**.
2. Enter **7** in the **Mean** box.
3. Enter **2** in the **Standard deviation** box.
4. Click **Input column** and enter **C1** in its box and press **Tab**.
5. Enter **C2** in the first **Optional storage** box.
6. Click **OK**.

Minitab displays the Example 6.4 Z value corresponding to a cumulative area of 0.10. Skip step 5 in either set of instructions to create the results shown in Figure 6.17 on page 230.

MG6.3 EVALUATING NORMALITY

Comparing Data Characteristics to Theoretical Properties

Use instructions in Sections MG3.1 through MG3.3 in the Chapter 3 Minitab Guide to compare data characteristics to theoretical properties.

Constructing the Normal Probability Plot

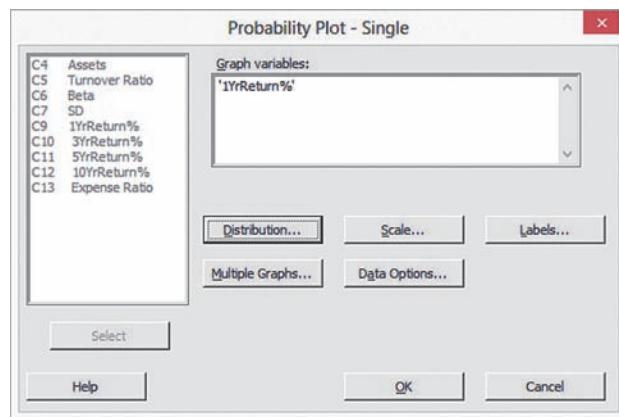
Use Probability Plot.

For example, to construct the normal probability plot for the one-year return percentage for the sample of 316 retirement funds shown in Figure 6.20 on page 236, open to the **Retirement Funds** worksheet. Select **Graph → Probability Plot** and:

1. In the Probability Plots dialog box, click **Single** and then click **OK**.

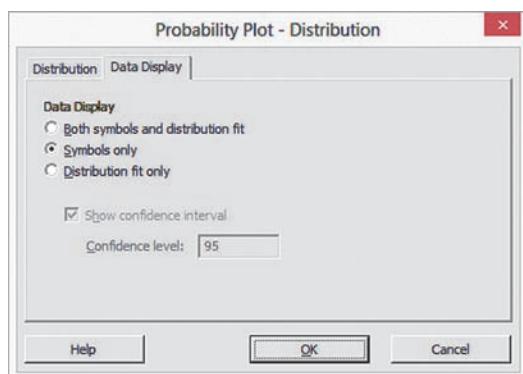
In the Probability Plot - Single dialog box (shown below):

2. Double-click **C9 1YrReturn%** in the variables list to add '**1YrReturn%**' to the **Graph variables** box.
3. Click **Distribution**.



In the Probability Plot - Distribution dialog box (shown below):

4. Click the **Distribution** tab and select **Normal** from the **Distribution** drop-down list.
5. Click the **Data Display** tab. Click **Symbols only**. If the **Show confidence interval** check box is not disabled (as shown below), clear this check box.
6. Click **OK**.



7. Back in the Probability Plot - Single dialog box, click **Scale**.
8. Click the **Gridlines** tab. Clear all check boxes and then click **OK**.
9. Back in the Probability Plot - Single dialog box, click **OK**.

MG6.4 The UNIFORM DISTRIBUTION

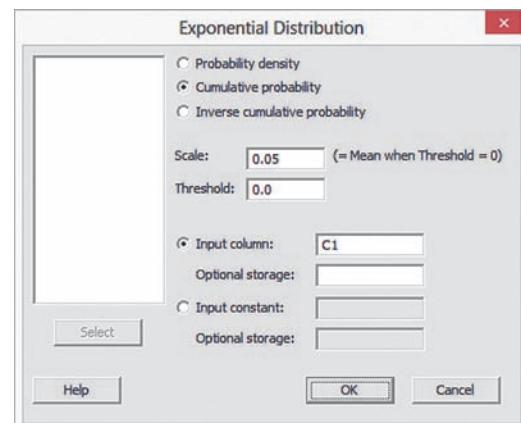
There are no Minitab instructions for this section

MG6.5 The EXPONENTIAL DISTRIBUTION

Use Exponential.

For example, to compute the exponential probability for the bank ATM customer arrival example on page 240, open to a new worksheet. Enter **X Value** as the name of column **C1** and enter **0.1** in the row 1 cell of column **C1**. Select **Calc → Probability Distributions → Exponential**. In the Exponential Distribution dialog box (shown below):

1. Click **Cumulative probability**.
2. Enter **0.05** in the **Scale** box. (Minitab defines scale as the mean time *between* arrivals, $1/\lambda = 1/20 = 0.05$, not the mean number of arrivals, $\lambda = 20$.)
3. Leave the **Threshold** value as **0.0**.
4. Click **Input column** and enter **C1** in its box.
5. Click **OK**.



CHAPTER

7

Sampling Distributions

CONTENTS

7.1 Sampling Distributions

7.2 Sampling Distribution of the Mean

VISUAL EXPLORATIONS:

[Exploring Sampling Distributions](#)

7.3 Sampling Distribution of the Proportion

7.4 Sampling from Finite Populations (*online*)

USING STATISTICS: Sampling Oxford Cereals, Revisited

CHAPTER 7 EXCEL GUIDE

CHAPTER 7 MINITAB GUIDE

OBJECTIVES

To learn about the concept of the sampling distribution

To compute probabilities related to the sample mean and the sample proportion

To understand the importance of the Central Limit Theorem

USING STATISTICS

Sampling Oxford Cereals

The automated production line at the Oxford Cereals main plant fills thousands of boxes of cereal during each shift. As the plant operations manager, you are responsible for monitoring the amount of cereal placed in each box. To be consistent with package labeling, boxes should contain a mean of 368 grams of cereal. Because of the speed of the process, the cereal weight varies from box to box, causing some boxes to be underfilled and others to be overfilled. If the automated process fails to work as intended, the mean weight in the boxes could vary too much from the label weight of 368 grams to be acceptable.

Because weighing every single box is too time-consuming, costly, and inefficient, you must take a sample of boxes. For each sample you select, you plan to weigh the individual boxes and calculate a sample mean. You need to determine the probability that such a sample mean could have been randomly selected from a population whose mean is 368 grams. Based on your analysis, you will have to decide whether to maintain, alter, or shut down the cereal-filling process.



Corbis

In Chapter 6, you used the normal distribution to study the distribution of video download times from the MyTVLab website. In this chapter, you need to make a decision about a cereal-filling process, based on the weights of a sample of cereal boxes packaged at Oxford Cereals. You will learn about sampling distributions and how to use them to solve business problems.

7.1 Sampling Distributions

In many applications, you want to make inferences that are based on statistics calculated from samples to estimate the values of population parameters. In the next two sections, you will learn about how the sample mean (a statistic) is used to estimate the population mean (a parameter) and how the sample proportion (a statistic) is used to estimate the population proportion (a parameter). Your main concern when making a statistical inference is reaching conclusions about a population, *not* about a sample. For example, a political pollster is interested in the sample results only as a way of estimating the actual proportion of the votes that each candidate will receive from the population of voters. Likewise, as plant operations manager for Oxford Cereals, you are only interested in using the mean weight calculated from a sample of cereal boxes to estimate the mean weight of a population of boxes.

In practice, you select a single random sample of a predetermined size from the population. Hypothetically, to use the sample statistic to estimate the population parameter, you could examine *every* possible sample of a given size that could occur. A **sampling distribution** is the distribution of the results if you actually selected all possible samples. The single result you obtain in practice is just one of the results in the sampling distribution.

7.2 Sampling Distribution of the Mean

In Chapter 3, several measures of central tendency, including the mean, median, and mode, were discussed. For several reasons, the mean is the most widely used measure of central tendency, and the sample mean is often used to estimate the population mean. The **sampling distribution of the mean** is the distribution of all possible sample means if you select all possible samples of a given size.

LEARN MORE

Learn more about the unbiased property of the sample mean in the **SHORT TAKES** for Chapter 7.

The Unbiased Property of the Sample Mean

The sample mean is **unbiased** because the mean of all the possible sample means (of a given sample size, n) is equal to the population mean, μ . A simple example concerning a population of four administrative assistants demonstrates this property. Each assistant is asked to apply the same set of updates to a human resources database. Table 7.1 presents the number of errors made by each of the administrative assistants. This population distribution is shown in Figure 7.1.

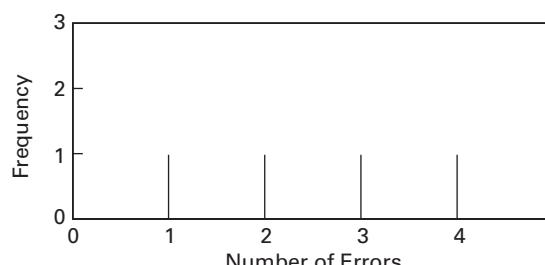
TABLE 7.1

Number of Errors Made by Each of Four Administrative Assistants

Administrative Assistant	Number of Errors
Ann	$X_1 = 3$
Bob	$X_2 = 2$
Carla	$X_3 = 1$
Dave	$X_4 = 4$

FIGURE 7.1

Number of errors made by a population of four administrative assistants



When you have data from a population, you compute the mean by using Equation (7.1), and you compute the population standard deviation, σ , by using Equation (7.2).

POPULATION MEAN

The population mean is the sum of the values in the population divided by the population size, N .

$$\mu = \frac{\sum_{i=1}^N X_i}{N} \quad (7.1)$$

POPULATION STANDARD DEVIATION

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}} \quad (7.2)$$

For the data of Table 7.1,

$$\mu = \frac{3 + 2 + 1 + 4}{4} = 2.5 \text{ errors}$$

and

$$\sigma = \sqrt{\frac{(3 - 2.5)^2 + (2 - 2.5)^2 + (1 - 2.5)^2 + (4 - 2.5)^2}{4}} = 1.12 \text{ errors}$$

If you select samples of two administrative assistants *with replacement* from this population, there are 16 possible samples ($N^n = 4^2 = 16$). Table 7.2 lists the 16 possible sample outcomes. If you average all 16 of these sample means, the mean of these values is equal to 2.5, which is also the mean of the population, μ ,

TABLE 7.2

All 16 Samples of
 $n = 2$ Administrative
Assistants from a
Population of $N = 4$
Administrative
Assistants When
Sampling with
Replacement

Sample	Administrative Assistants	Sample Outcomes	Sample Mean
1	Ann, Ann	3, 3	$\bar{X}_1 = 3$
2	Ann, Bob	3, 2	$\bar{X}_2 = 2.5$
3	Ann, Carla	3, 1	$\bar{X}_3 = 2$
4	Ann, Dave	3, 4	$\bar{X}_4 = 3.5$
5	Bob, Ann	2, 3	$\bar{X}_5 = 2.5$
6	Bob, Bob	2, 2	$\bar{X}_6 = 2$
7	Bob, Carla	2, 1	$\bar{X}_7 = 1.5$
8	Bob, Dave	2, 4	$\bar{X}_8 = 3$
9	Carla, Ann	1, 3	$\bar{X}_9 = 2$
10	Carla, Bob	1, 2	$\bar{X}_{10} = 1.5$
11	Carla, Carla	1, 1	$\bar{X}_{11} = 1$
12	Carla, Dave	1, 4	$\bar{X}_{12} = 2.5$
13	Dave, Ann	4, 3	$\bar{X}_{13} = 3.5$
14	Dave, Bob	4, 2	$\bar{X}_{14} = 3$
15	Dave, Carla	4, 1	$\bar{X}_{15} = 2.5$
16	Dave, Dave	4, 4	$\bar{X}_{16} = 4$
			$\mu_{\bar{X}} = 2.5$

Because the mean of the 16 sample means is equal to the population mean, the sample mean is an unbiased estimator of the population mean. Therefore, although you do not know how close the sample mean of any particular sample selected is to the population mean, you are assured that the mean of all the possible sample means that could have been selected is equal to the population mean.

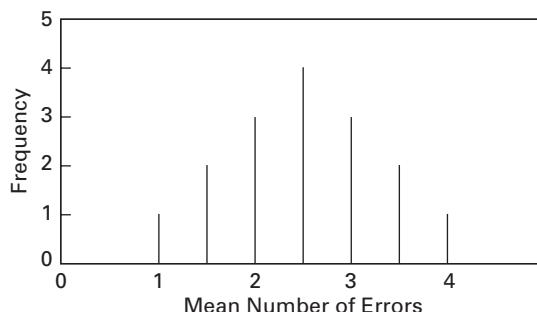
Standard Error of the Mean

Figure 7.2 illustrates the variation in the sample means when selecting all 16 possible samples.

FIGURE 7.2

Sampling distribution of the mean, based on all possible samples containing two administrative assistants

Source: Data are from Table 7.2.



In this small example, although the sample means vary from sample to sample, depending on which two administrative assistants are selected, the sample means do not vary as much as the individual values in the population. That the sample means are less variable than the individual values in the population follows directly from the fact that each sample mean averages together all the values in the sample. A population consists of individual outcomes that can take on a wide range of values, from extremely small to extremely large. However, if a sample contains an extreme value, although this value will have an effect on the sample mean, the effect is reduced because the value is averaged with all the other values in the sample. As the sample size increases, the effect of a single extreme value becomes smaller because it is averaged with more values.

The value of the standard deviation of all possible sample means, called the **standard error of the mean**, expresses how the sample means vary from sample to sample. As the sample size increases, the standard error of the mean decreases by a factor equal to the square root of the sample size. Equation (7.3) defines the standard error of the mean when sampling *with replacement* or sampling *without replacement* from large or infinite populations.

Student Tip

Remember, the standard error of the mean measures variation among the means not the individual values.

STANDARD ERROR OF THE MEAN

The standard error of the mean, $\sigma_{\bar{X}}$, is equal to the standard deviation in the population, σ , divided by the square root of the sample size, n .

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (7.3)$$

Example 7.1 computes the standard error of the mean when the sample selected without replacement contains less than 5% of the entire population.

EXAMPLE 7.1**Computing the Standard Error of the Mean**

Returning to the cereal-filling process described in the Using Statistics scenario on page 250, if you randomly select a sample of 25 boxes without replacement from the thousands of boxes filled during a shift, the sample contains much less than 5% of the population. Given that the standard deviation of the cereal-filling process is 15 grams, compute the standard error of the mean.

SOLUTION Using Equation (7.3) with $n = 25$ and $\sigma = 15$ the standard error of the mean is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3$$

The variation in the sample means for samples of $n = 25$ is much less than the variation in the individual boxes of cereal (i.e., $\sigma_{\bar{X}} = 3$, while $\sigma = 15$).

Sampling from Normally Distributed Populations

Now that the concept of a sampling distribution has been introduced and the standard error of the mean has been defined, what distribution will the sample mean, \bar{X} , follow? If you are sampling from a population that is normally distributed with mean μ and standard deviation σ , then regardless of the sample size, n , the sampling distribution of the mean is normally distributed, with mean $\mu_{\bar{X}} = \mu$, and standard error of the mean $\sigma_{\bar{X}} = \sigma/\sqrt{n}$.

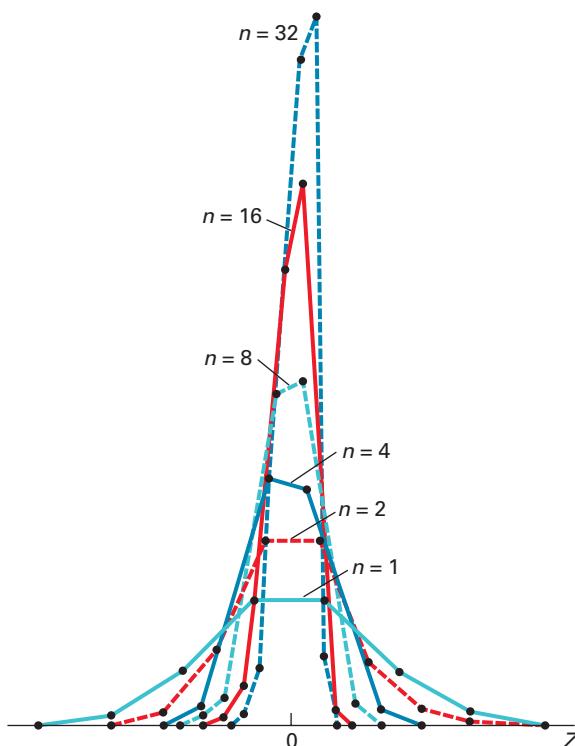
In the simplest case, if you take samples of size $n = 1$, each possible sample mean is a single value from the population because

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1}{1} = X_1$$

Therefore, if the population is normally distributed, with mean μ and standard deviation σ , the sampling distribution \bar{X} for samples of $n = 1$ must also follow the normal distribution, with mean $\mu_{\bar{X}} = \mu$ and standard error of the mean $\sigma_{\bar{X}} = \sigma/\sqrt{1} = \sigma$. In addition, as the sample size increases, the sampling distribution of the mean still follows a normal distribution, with $\mu_{\bar{X}} = \mu$, but the standard error of the mean decreases so that a larger proportion of sample means are closer to the population mean. Figure 7.3 illustrates this reduction in variability.

FIGURE 7.3

Sampling distributions of the mean from 500 samples of sizes $n = 1, 2, 4, 8, 16$, and 32 selected from a normal population



¹Remember that “only” 500 samples out of an infinite number of samples have been selected, so that the sampling distributions shown are only approximations of the population distribution.

Note that 500 samples of size 1, 2, 4, 8, 16, and 32 were randomly selected from a normally distributed population. From the polygons in Figure 7.3, you can see that, although the sampling distribution of the mean is approximately¹ normal for each sample size, the sample means are distributed more tightly around the population mean as the sample size increases.

To further examine the concept of the sampling distribution of the mean, consider the Using Statistics scenario described on page 250. The packaging equipment that is filling 368-gram boxes of cereal is set so that the amount of cereal in a box is normally distributed, with a mean of 368 grams. From past experience, you know the population standard deviation for this filling process is 15 grams.

If you randomly select a sample of 25 boxes from the many thousands that are filled in a day and the mean weight is computed for this sample, what type of result could you expect? For example, do you think that the sample mean could be 368 grams? 200 grams? 365 grams?

The sample acts as a miniature representation of the population, so if the values in the population are normally distributed, the values in the sample should be approximately normally distributed. Thus, if the population mean is 368 grams, the sample mean has a good chance of being close to 368 grams.

How can you determine the probability that the sample of 25 boxes will have a mean below 365 grams? From the normal distribution (Section 6.2), you know that you can find the area below any value X by converting to standardized Z values:

$$Z = \frac{X - \mu}{\sigma}$$

In the examples in Section 6.2, you studied how any single value, X , differs from the population mean. Now, in this example, you want to study how a sample mean, \bar{X} , differs from the population mean. Substituting \bar{X} for X , $\mu_{\bar{X}}$ for μ , and $\sigma_{\bar{X}}$ for σ in the equation above results in Equation (7.4).

FINDING Z FOR THE SAMPLING DISTRIBUTION OF THE MEAN

The Z value is equal to the difference between the sample mean, \bar{X} , and the population mean, μ , divided by the standard error of the mean, $\sigma_{\bar{X}}$.

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad (7.4)$$

To find the area below 365 grams, from Equation (7.4),

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{365 - 368}{\frac{15}{\sqrt{25}}} = \frac{-3}{3} = -1.00$$

The area corresponding to $Z = -1.00$ in Table E.2 is 0.1587. Therefore, 15.87% of all the possible samples of 25 boxes have a sample mean below 365 grams.

The preceding statement is not the same as saying that a certain percentage of *individual* boxes will contain less than 365 grams of cereal. You compute that percentage as follows:

$$Z = \frac{X - \mu}{\sigma} = \frac{365 - 368}{15} = \frac{-3}{15} = -0.20$$

The area corresponding to $Z = -0.20$ in Table E.2 is 0.4207. Therefore, 42.07% of the *individual* boxes are expected to contain less than 365 grams. Comparing these results, you see that many more *individual boxes* than *sample means* are below 365 grams. This result is explained by the fact that each sample consists of 25 different values, some small and some large. The averaging

process dilutes the importance of any individual value, particularly when the sample size is large. Therefore, the chance that the sample mean of 25 boxes is very different from the population mean is less than the chance that a *single* box is very different from the population mean.

Examples 7.2 and 7.3 show how these results are affected by using different sample sizes.

EXAMPLE 7.2

The Effect of Sample Size, n , on the Computation of $\sigma_{\bar{X}}$

How is the standard error of the mean affected by increasing the sample size from 25 to 100 boxes?

SOLUTION If $n = 100$ boxes, then using Equation (7.3) on page 253,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$$

The fourfold increase in the sample size from 25 to 100 reduces the standard error of the mean by half—from 3 grams to 1.5 grams. This demonstrates that taking a larger sample results in less variability in the sample means from sample to sample.

EXAMPLE 7.3

The Effect of Sample Size, n , on the Clustering of Means in the Sampling Distribution

If you select a sample of 100 boxes, what is the probability that the sample mean is below 365 grams?

SOLUTION Using Equation (7.4) on page 255,

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{365 - 368}{\frac{15}{\sqrt{100}}} = \frac{-3}{1.5} = -2.00$$

From Table E.2, the area less than $Z = -2.00$ is 0.0228. Therefore, 2.28% of the samples of 100 boxes have means below 365 grams, as compared with 15.87% for samples of 25 boxes.

Sometimes you need to find the interval that contains a specific proportion of the sample means. To do so, you determine a distance below and above the population mean containing a specific area of the normal curve. From Equation (7.4) on page 255,

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Solving for \bar{X} results in Equation (7.5).

FINDING \bar{X} FOR THE SAMPLING DISTRIBUTION OF THE MEAN

$$\bar{X} = \mu + Z \frac{\sigma}{\sqrt{n}} \quad (7.5)$$

Example 7.4 illustrates the use of Equation (7.5).

EXAMPLE 7.4

Determining the Interval That Includes a Fixed Proportion of the Sample Means

In the cereal-filling example, find an interval symmetrically distributed around the population mean that will include 95% of the sample means, based on samples of 25 boxes.

SOLUTION If 95% of the sample means are in the interval, then 5% are outside the interval. Divide the 5% into two equal parts of 2.5%. The value of Z in Table E.2 corresponding to an area of 0.0250 in the lower tail of the normal curve is -1.96 , and the value of Z corresponding to a cumulative area of 0.9750 (i.e., 0.0250 in the upper tail of the normal curve) is $+1.96$.

The lower value of \bar{X} (called \bar{X}_L) and the upper value of \bar{X} (called \bar{X}_U) are found by using Equation (7.5):

$$\bar{X}_L = 368 + (-1.96) \frac{15}{\sqrt{25}} = 368 - 5.88 = 362.12$$

$$\bar{X}_U = 368 + (1.96) \frac{15}{\sqrt{25}} = 368 + 5.88 = 373.88$$

Therefore, 95% of all sample means, based on samples of 25 boxes, are between 362.12 and 373.88 grams.

Sampling from Non-normally Distributed Populations—The Central Limit Theorem

So far in this section, only the sampling distribution of the mean for a normally distributed population has been considered. However, for many analyses, you will either be able to know that the population is not normally distributed or conclude that it would be unrealistic to assume that the population is normally distributed. An important theorem in statistics, the **Central Limit Theorem**, deals with these situations.

THE CENTRAL LIMIT THEOREM

As the sample size (the number of values in each sample) gets *large enough*, the sampling distribution of the mean is approximately normally distributed. This is true regardless of the shape of the distribution of the individual values in the population.

What sample size is *large enough*? A great deal of statistical research has gone into this issue. As a general rule, statisticians have found that for many population distributions, when the sample size is at least 30, the sampling distribution of the mean is approximately normal. However, you can apply the Central Limit Theorem for even smaller sample sizes if the population distribution is approximately bell-shaped. In the case in which the distribution of a variable is extremely skewed or has more than one mode, you may need sample sizes larger than 30 to ensure normality in the sampling distribution of the mean.

Figure 7.4 shows the sampling distributions from three different continuous distributions (normal, uniform, and exponential) for varying sample sizes ($n = 2, 5$, and 30) and illustrates the application of the Central Limit Theorem to these different populations. In each of the panels, because the sample mean is an unbiased estimator of the population mean, the mean of any sampling distribution is always equal to the mean of the population.

Figure 7.4 Panel A shows the sampling distribution of the mean selected from a normal population. As mentioned earlier in this section, when the population is normally distributed, the sampling distribution of the mean is normally distributed for *any* sample size. [You can measure the variability by using the standard error of the mean, Equation (7.3), on page 253.] Observe that for each of the sample sizes illustrated, the shape of the sampling distribution is a bell-shaped normal distribution. As the sample size increases from 2 to 30, you can see that the variation becomes smaller, resulting in a distribution that has less variation around the mean.

FIGURE 7.4

Sampling distribution of the mean for different populations for samples of $n = 2, 5$ and 30

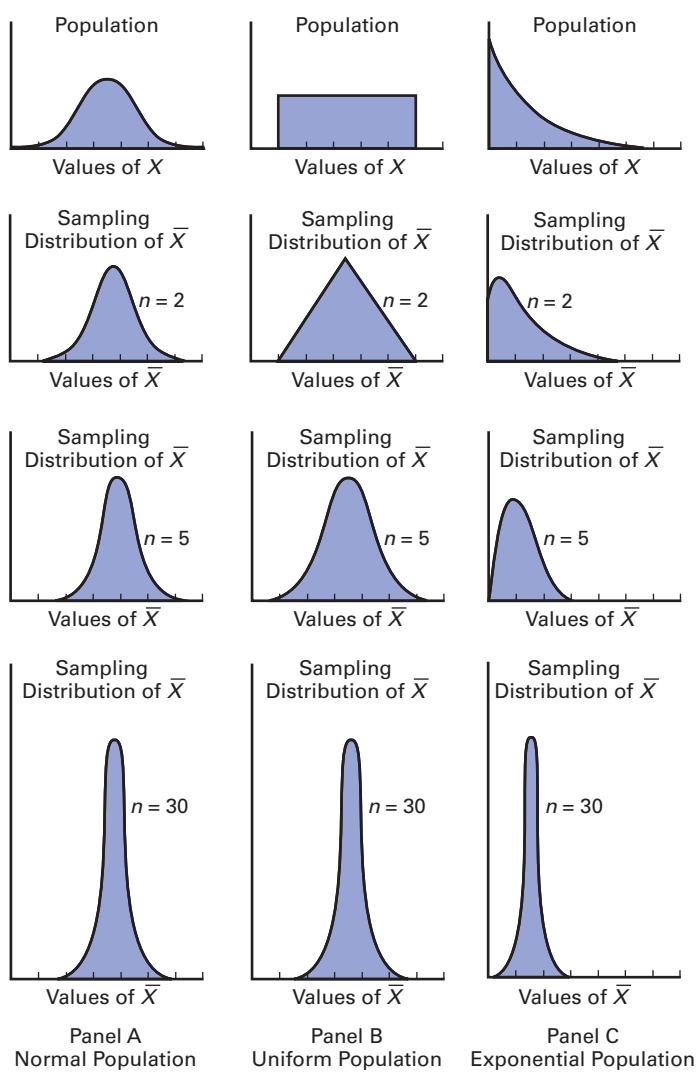


Figure 7.4 Panel B depicts the sampling distribution from a population with a uniform (or rectangular) distribution (see Section 6.4). When samples of size $n = 2$ are selected, there is a peaking, or *central limiting*, effect already working, resulting in a sampling distribution that looks like a triangle. For $n = 5$, the sampling distribution is bell-shaped and approximately normal. When $n = 30$, the sampling distribution looks very similar to a normal distribution. In general, the larger the sample size, the more closely the sampling distribution will follow a normal distribution. As with all other cases, the mean of each sampling distribution is equal to the mean of the population, and the variability decreases as the sample size increases.

Figure 7.4 Panel C presents an exponential distribution (see Section 6.5). This population is extremely right-skewed. When $n = 2$, the sampling distribution is still highly right-skewed but less so than the distribution of the population. For $n = 5$, the sampling distribution is slightly right-skewed. When $n = 30$, the sampling distribution looks approximately normal. Again, the mean of each sampling distribution is equal to the mean of the population, and the variability decreases as the sample size increases.

Using the results from the normal, uniform, and exponential distributions, you can reach the following conclusions regarding the Central Limit Theorem:

- For most distributions, regardless of the shape of the population, the sampling distribution of the mean is approximately normally distributed if samples of at least size 30 are selected.
- If the distribution of the population is fairly symmetrical, the sampling distribution of the mean is approximately normal for samples as small as size 5.
- If the population is normally distributed, the sampling distribution of the mean is normally distributed, regardless of the sample size.

The Central Limit Theorem is of crucial importance in using statistical inference to reach conclusions about a population. It allows you to make inferences about the population mean without having to know the specific shape of the population distribution. Example 7.5 illustrates a sampling distribution for a skewed population.

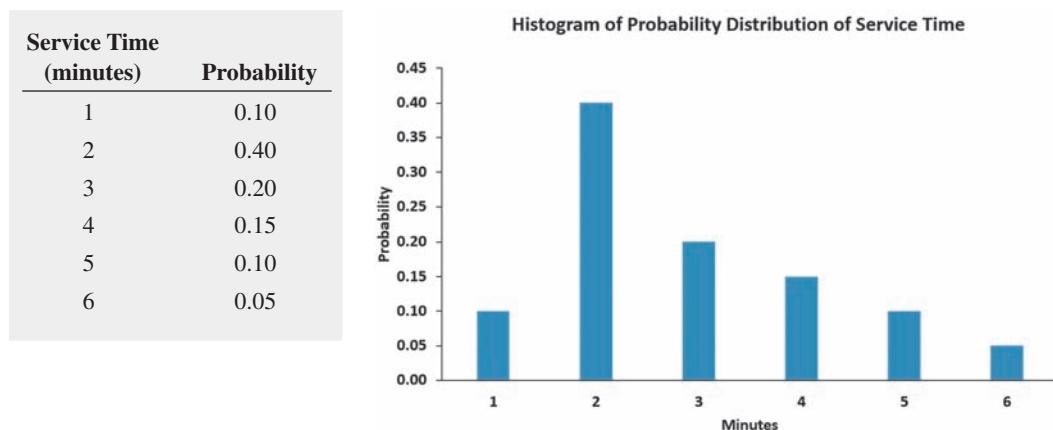
EXAMPLE 7.5

Constructing a Sampling Distribution for a Skewed Population

Figure 7.5 shows the distribution of the time it takes to fill orders at a fast-food chain drive-through lane. Note that the probability distribution table is unlike Table 7.1 (page 251), which presents a population in which each value is equally likely to occur.

FIGURE 7.5

Probability distribution of the service time (in minutes) at a fast-food chain drive-through lane and a histogram of that probability distribution



Using Equation 5.1 on page 187, the population mean is computed as 2.9 minutes. Using Equation (5.3) on page 188, the population standard deviation is computed as 1.34. Select 100 samples of $n = 2$, $n = 15$, and $n = 30$. What conclusions can you reach about the sampling distribution of the service time (in minutes) at the fast-food chain drive-through lane?

SOLUTION Table 7.3 represents the mean service time (in minutes) at the fast-food chain drive-through lane for 100 different random samples of $n = 2$. The mean of these 100 sample means is 2.825 minutes, and the standard error of the mean is 0.883.

TABLE 7.3

Mean Service Times (in minutes) at a Fast-Food Chain Drive-Through Lane for 100 Different Random Samples of $n = 2$

3.5	2.5	3	3.5	4	3	2.5	2	2	2.5
3	3	2.5	2.5	2	2.5	2.5	2	3.5	1.5
2	3	2.5	3	3	2	3.5	3.5	2.5	2
4.5	3.5	4	2	2	4	3.5	2.5	2.5	3.5
3.5	3.5	2	1.5	2.5	2	3.5	3.5	2.5	2.5
2.5	3	3	3.5	2	3.5	2	1.5	5.5	2.5
3.5	3	3	2	1.5	3	2.5	2.5	2.5	2.5
3.5	1.5	6	2	1.5	2.5	3.5	2	3.5	5
2.5	3.5	4.5	3.5	3.5	2	4	2	3	3
4.5	1.5	2.5	2	2.5	2.5	2	2	2	4

Table 7.4 represents the mean service time (in minutes) at the fast-food chain drive-through lane for 100 different random samples of $n = 15$. The mean of these 100 sample means is 2.9313 minutes, and the standard error of the mean is 0.3458.

Table 7.5 represents the mean service time (in minutes) at the fast-food chain drive-through lane for 100 different random samples of $n = 30$. The mean of these 100 sample means is 2.9527 minutes, and the standard error of the mean is 0.2701.

(continued)

TABLE 7.4Mean Service Times (in minutes) at a Fast-Food Chain Drive-Through Lane for 100 Different Random Samples of $n = 15$

3.5333	2.8667	3.1333	3.6000	2.5333	2.8000	2.8667	3.1333	3.2667	3.3333
3.0000	3.3333	2.7333	2.6000	2.8667	3.0667	2.1333	2.5333	2.8000	3.1333
2.8000	2.7333	2.6000	3.1333	2.8667	3.4667	2.9333	2.8000	2.2000	3.0000
2.9333	2.6000	2.6000	3.1333	3.1333	3.1333	2.5333	3.0667	3.9333	2.8000
3.0000	2.7333	2.6000	2.4667	3.2000	2.4667	3.2000	2.9333	2.8667	3.4667
2.6667	3.0000	3.1333	3.1333	2.7333	2.7333	3.3333	3.4000	3.2000	3.0000
3.2000	3.0000	2.6000	2.9333	3.0667	2.8667	2.2667	2.5333	2.7333	2.2667
2.8000	2.8000	2.6000	3.1333	2.9333	3.0667	3.6667	2.6667	2.8667	2.6667
3.0000	3.4000	2.7333	3.6000	2.6000	2.7333	3.3333	2.6000	2.8667	2.8000
3.7333	2.9333	3.0667	2.6667	2.8667	2.2667	2.7333	2.8667	3.5333	3.2000

TABLE 7.5Mean Service Times (in minutes) at a Fast-Food Chain Drive-Through Lane for 100 Different Random Samples of $n = 30$

3.0000	3.3667	3.0000	3.1333	2.8667	2.8333	3.2667	2.9000	2.7000	3.2000
3.2333	2.7667	3.2333	2.8000	3.4000	3.0333	2.8667	3.0000	3.1333	3.4000
2.3000	3.0000	3.0667	2.9667	3.0333	2.4000	2.8667	2.8000	2.5000	2.7000
2.7000	2.9000	2.8333	3.3000	3.1333	2.8667	2.6667	2.6000	3.2333	2.8667
2.7667	2.9333	2.5667	2.5333	3.0333	3.2333	3.0667	2.9667	2.4000	3.3000
2.8000	3.0667	3.2000	2.9667	2.9667	3.2333	3.3667	2.9000	3.0333	3.1333
3.3333	2.8667	2.8333	3.0667	3.3667	3.0667	3.0667	3.2000	3.1667	3.3667
3.0333	3.1667	2.4667	3.0000	2.6333	2.6667	2.9667	3.1333	2.8000	2.8333
2.9333	2.7000	3.0333	2.7333	2.6667	2.6333	3.1333	3.0667	2.5333	3.3333
3.1000	2.5667	2.9000	3.9333	2.9000	2.7000	2.7333	2.8000	2.6667	2.8333

Figure 7.6 Panels A through C show histograms of the mean service time (in minutes) at the fast-food chain drive-through lane for the three sets of 100 different random samples shown in Tables 7.3 through 7.5. Panel A, the histogram for the mean service time for 100 different random samples of $n = 2$, shows a skewed distribution, but a distribution that is not as skewed as the population distribution of service times shown in Figure 7.5.

Panel B, the histogram for the mean service time for 100 different random samples of $n = 15$, shows a somewhat symmetrical distribution that contains a concentration of values in the center of the distribution. Panel C, the histogram for the mean service time for 100 different

FIGURE 7.6

Histograms of the mean service time (in minutes) at the fast-food chain drive-through lane of 100 different random samples of $n = 2$ (Panel A, left) and 100 different random samples of $n = 15$ (Panel B, right)

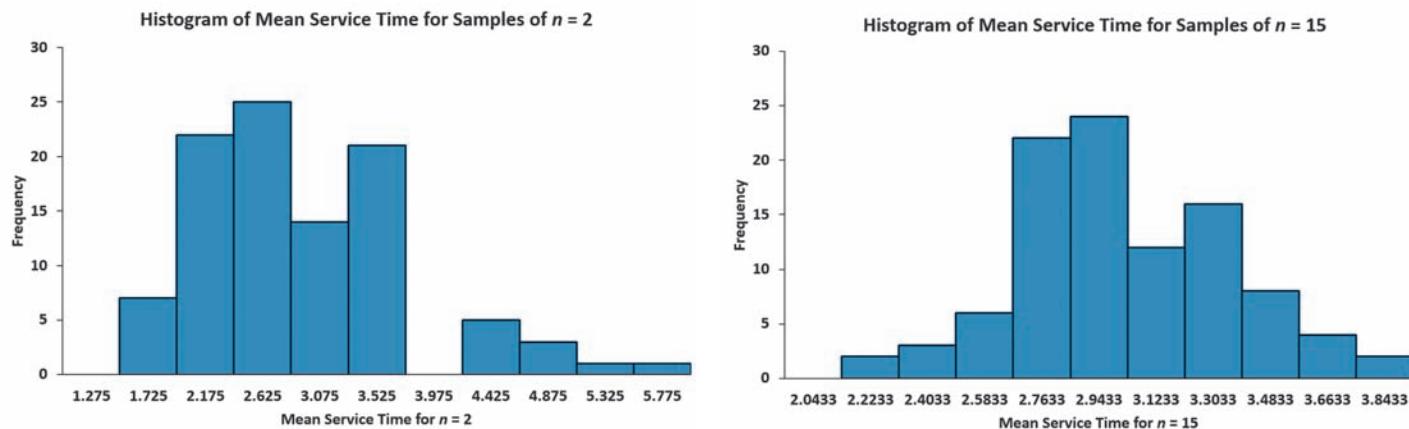
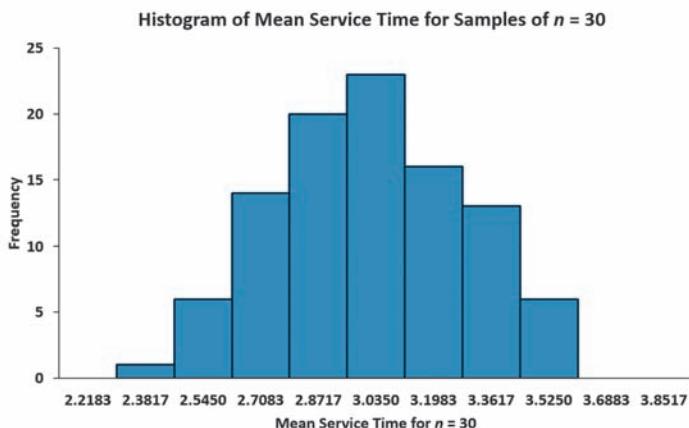


FIGURE 7.6 (continued)

Panel C: Histogram of the mean service time (in minutes) at the fast-food chain drive-through lane of 100 different random samples of $n = 30$



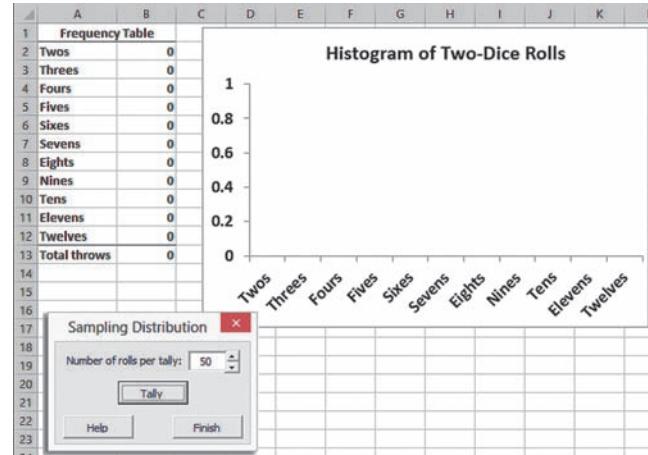
random samples of $n = 30$, shows a distribution that appears to be approximately bell-shaped with a concentration of values in the center of the distribution. The progression of the histograms from a skewed population towards a bell-shaped distribution as the sample size increases is consistent with the Central Limit Theorem.

VISUAL EXPLORATIONS

Exploring Sampling Distributions

Open the **VE-Sampling Distribution add-in workbook** to observe the effects of simulated rolls on the frequency distribution of the sum of two dice. (For Excel technical requirements, review Appendix Section D.4) When this workbook opens properly, it adds a Sampling Distribution menu to the Add-ins tab (Apple menu in Excel 2011).

To observe the effects of simulated throws on the frequency distribution of the sum of the two dice, select **Sampling Distribution** → **Two Dice Simulation**. In the Sampling Distribution dialog box, enter the **Number of rolls per tally** and click **Tally**. Click **Finish** when done.



Problems for Section 7.2

LEARNING THE BASICS

- 7.1** Given a normal distribution with $\mu = 100$ and $\sigma = 10$, if you select a sample of $n = 25$, what is the probability that \bar{X} is
- less than 95?
 - between 95 and 97.5?
 - above 102.2?
 - There is a 65% chance that \bar{X} is above what value?

- 7.2** Given a normal distribution with $\mu = 50$ and $\sigma = 5$, if you select a sample of $n = 100$, what is the probability that \bar{X} is
- less than 47?
 - between 47 and 49.5?
 - above 51.1?
 - There is a 35% chance that \bar{X} is above what value?

APPLYING THE CONCEPTS

7.3 For each of the following three populations, indicate what the sampling distribution for samples of 25 would consist of:

- Customer receipts for a supermarket for a year.
- Insurance payouts in a particular geographical area in a year.
- Call center logs of inbound calls tracking handling time for a credit card company during the year.

7.4 The following data represent the number of days absent per year in a population of six employees of a small company:

1 3 6 7 9 10

- Assuming that you sample without replacement, select all possible samples of $n = 2$ and construct the sampling distribution of the mean. Compute the mean of all the sample means and also compute the population mean. Are they equal? What is this property called?
- Repeat (a) for all possible samples of $n = 3$.
- Compare the shape of the sampling distribution of the mean in (a) and (b). Which sampling distribution has less variability? Why?
- Assuming that you sample with replacement, repeat (a) through (c) and compare the results. Which sampling distributions have the least variability—those in (a) or (b)? Why?

7.5 The diameter of a brand of tennis balls is approximately normally distributed, with a mean of 2.63 inches and a standard deviation of 0.03 inch. If you select a random sample of 9 tennis balls,

- what is the sampling distribution of the mean?
- what is the probability that the sample mean is less than 2.61 inches?
- what is the probability that the sample mean is between 2.62 and 2.64 inches?
- The probability is 60% that the sample mean will be between what two values symmetrically distributed around the population mean?

7.6 The U.S. Census Bureau announced that the median sales price of new houses sold in 2012 was \$245,200, and the mean sales price was \$291,200 (www.census.gov/newhomesales, April 1, 2013). Assume that the standard deviation of the prices is \$90,000.

- If you select samples of $n = 4$, describe the shape of the sampling distribution of \bar{X} .
- If you select samples of $n = 100$, describe the shape of the sampling distribution of \bar{X} .
- If you select a random sample of $n = 100$, what is the probability that the sample mean will be less than \$315,000?
- If you select a random sample of $n = 100$, what is the probability that the sample mean will be between \$295,000 and \$310,000?

7.7 According to a mashable.com post, time spent on Tumblr, a microblogging platform and social networking website, has a mean of 14 minutes per visit. (Source: on.mash.to/1757wfE.) Assume that time spent on Tumblr per visit is normally distributed and that the standard deviation is 4 minutes. If you select a random sample of 25 visits,

- what is the probability that the sample mean is between 13.6 and 14.4 minutes?
- what is the probability that the sample mean is between 13 and 14 minutes?
- If you select a random sample of 100 visits, what is the probability that the sample mean is between 13.6 and 14.4 minutes?
- Explain the difference in the results of (a) and (c).



7.8 Today, full-time college students report spending a mean of 27 hours per week on academic activities, both inside and outside the classroom. (Source: “A Challenge to College Students for 2013: Don’t Waste Your 6,570,” *Huffington Post*, January 29, 2013, huff.to/13dNtuT.) Assume the standard deviation of time spent on academic activities is 4 hours. If you select a random sample of 16 full-time college students,

- what is the probability that the mean time spent on academic activities is at least 26 hours per week?
- there is an 85% chance that the sample mean is less than how many hours per week?
- What assumption must you make in order to solve (a) and (b)?
- If you select a random sample of 64 full-time college students, there is an 85% chance that the sample mean is less than how many hours per week?

7.3 Sampling Distribution of the Proportion

Student Tip

Do not confuse this use of the Greek letter pi, π , to represent the population proportion with the mathematical constant that is the ratio of the circumference to a diameter of a circle—approximately 3.14159—which is also known by the same Greek letter.

Consider a categorical variable that has only two categories, such as the customer prefers your brand or the customer prefers the competitor’s brand. You are interested in the proportion of items belonging to one of the categories—for example, the proportion of customers that prefer your brand. The population proportion, represented by π , is the proportion of items in the entire population with the characteristic of interest. The sample proportion, represented by p , is the proportion of items in the sample with the characteristic of interest. The sample proportion, a statistic, is used to estimate the population proportion, a parameter. To calculate the sample proportion, you assign one of two possible values, 1 or 0, to represent the presence or absence of the characteristic. You then sum all the 1 and 0 values and divide by n , the sample size. For example, if, in a sample of five customers, three preferred your brand and two did not, you have three 1s and two 0s. Summing the three 1s and two 0s and dividing by the sample size of 5 results in a sample proportion of 0.60.

SAMPLE PROPORTION

$$p = \frac{X}{n} = \frac{\text{Number of items having the characteristic of interest}}{\text{Sample size}} \quad (7.6)$$

Student Tip

Remember that the sample proportion cannot be negative and also cannot be greater than 1.0.

The sample proportion, p , will be between 0 and 1. If all items have the characteristic, you assign each a score of 1, and p is equal to 1. If half the items have the characteristic, you assign half a score of 1 and assign the other half a score of 0, and p is equal to 0.5. If none of the items have the characteristic, you assign each a score of 0, and p is equal to 0.

In Section 7.2, you learned that the sample mean, \bar{X} , is an unbiased estimator of the population mean, μ . Similarly, the statistic p is an unbiased estimator of the population proportion, π .

By analogy to the sampling distribution of the mean, whose standard error is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, the **standard error of the proportion**, σ_p , is given in Equation (7.7).

STANDARD ERROR OF THE PROPORTION

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}} \quad (7.7)$$

The **sampling distribution of the proportion** follows the binomial distribution, as discussed in Section 5.3, when sampling with replacement (or without replacement from extremely large populations). However, you can use the normal distribution to approximate the binomial distribution when $n\pi$ and $n(1 - \pi)$ are each at least 5. In most cases in which inferences are made about the population proportion, the sample size is substantial enough to meet the conditions for using the normal approximation (see reference 1). Therefore, in many instances, you can use the normal distribution to estimate the sampling distribution of the proportion.

Substituting p for \bar{X} , π for μ , and $\sqrt{\frac{\pi(1 - \pi)}{n}}$ for $\frac{\sigma}{\sqrt{n}}$ in Equation (7.4) on page 255 results in Equation (7.8).

FINDING Z FOR THE SAMPLING DISTRIBUTION OF THE PROPORTION

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \quad (7.8)$$

To illustrate the sampling distribution of the proportion, a recent survey (“Can You Stop Thinking About Work on Your Vacation?” *USA Today Snapshots*, October 5, 2011, p. 1A) reported that 32% of adults are unable to stop thinking about work while on vacation. Suppose that you select a random sample of 200 vacationers who have booked tours from a certain tour company, and you want to determine the probability that more than 40% of the vacationers are unable to stop thinking about work while on vacation. Because $n\pi = 200(0.32) = 64 > 5$ and $n(1 - \pi) = 200(1 - 0.32) = 136 > 5$, the sample size is large enough to assume that the sampling distribution of the proportion is approximately normally distributed. Then, using

the survey percentage of 32% as the population proportion, you can calculate the probability that more than 40% of the vacationers are unable to stop thinking about work while on vacation by using Equation (7.8):

$$\begin{aligned} Z &= \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \\ &= \frac{0.40 - 0.32}{\sqrt{\frac{(0.32)(0.68)}{200}}} = \frac{0.08}{\sqrt{\frac{0.2176}{200}}} = \frac{0.08}{0.0330} \\ &= 2.42 \end{aligned}$$

Using Table E.2, the area under the normal curve greater than 2.42 is 0.0078. Therefore, if the population proportion is 0.32, the probability is 0.78% that more than 40% of the 200 vacationers in the sample will be unable to stop thinking about work while on vacation.

Problems for Section 7.3

LEARNING THE BASICS

7.9 In a random sample of 64 people, 48 are classified as “successful.”

- Determine the sample proportion, p , of “successful” people.
- If the population proportion is 0.70, determine the standard error of the proportion.

7.10 A random sample of 50 households was selected for a phone (landline and cellphone) survey. The key question asked was, “Do you or any member of your household own an Apple product (iPhone, iPod, iPad, or Mac computer)?” Of the 50 respondents, 20 said yes and 30 said no.

- Determine the sample proportion, p , of households that own an Apple product.
- If the population proportion is 0.45, determine the standard error of the proportion.

7.11 The following data represent the responses (Y for yes and N for no) from a sample of 40 college students to the question “Do you currently own shares in any stocks?”

N N Y N N Y N Y N Y N N Y N Y N N N Y
N Y N N N N Y N N Y Y N N N Y N N Y N N

- Determine the sample proportion, p , of college students who own shares of stock.
- If the population proportion is 0.30, determine the standard error of the proportion.

APPLYING THE CONCEPTS

SELF Test **7.12** A political pollster is conducting an analysis of sample results in order to make predictions on election night. Assuming a two-candidate election, if a specific candidate receives at least 55% of the vote in the sample, that candidate will be forecast as the winner of the election. If you select a random sample of 100 voters, what is the probability that a candidate will be forecast as the winner when

- the population percentage of her vote is 50.1%?
- the population percentage of her vote is 60%?
- the population percentage of her vote is 49% (and she will actually lose the election)?
- If the sample size is increased to 400, what are your answers to (a) through (c)? Discuss.

7.13 You plan to conduct a marketing experiment in which students are to taste one of two different brands of soft drink. Their task is to correctly identify the brand tasted. You select a random sample of 200 students and assume that the students have no ability to distinguish between the two brands. (Hint: If an individual has no ability to distinguish between the two soft drinks, then the two brands are equally likely to be selected.)

- What is the probability that the sample will have between 50% and 60% of the identifications correct?
- The probability is 90% that the sample percentage is contained within what symmetrical limits of the population percentage?
- What is the probability that the sample percentage of correct identifications is greater than 65%?
- Which is more likely to occur—more than 60% correct identifications in the sample of 200 or more than 55% correct identifications in a sample of 1,000? Explain.

7.14 Accenture’s *Defining Success* global research study found that the majority of today’s working women would prefer a better work–life balance to an increased salary. One of the most important contributors to work–life balance identified by the survey was “flexibility,” with 80% of women saying that having a flexible work schedule is either very important or extremely important to their career success. (Source: bit.ly/17IM8gq.) Suppose you select a sample of 100 working women.

- What is the probability that in the sample fewer than 85% say that having a flexible work schedule is either very important or extremely important to their career success?
- What is the probability that in the sample between 75% and 85% say that having a flexible work schedule is either very important or extremely important to their career success?

- c. What is the probability that in the sample more than 82% say that having a flexible work schedule is either very important or extremely important to their career success?
- d. If a sample of 400 is taken, how does this change your answers to (a) through (c)?

7.15 The goal of corporate sustainability is to manage the environmental, economic, and social effects of a corporation's operations so it is profitable over the long term while acting in a responsible manner toward society. A Hill + Knowlton Strategies survey found that 57% of U.S. respondents are more likely to buy stock in a U.S. corporation, or shop at its stores, if it is making an effort to publicly talk about how it is becoming more sustainable. (Source: "Sustainability," bit.ly/10A2Snl.) Suppose you select a sample of 100 U.S. respondents.

- a. What is the probability that in the sample, fewer than 57% are more likely to buy stock in a U.S. corporation, or shop at its stores, if it is making an effort to publicly talk about how it is becoming more sustainable?
- b. What is the probability that in the sample, between 52% and 62% are more likely to buy stock in a U.S. corporation, or shop at its stores, if it is making an effort to publicly talk about how it is becoming more sustainable?
- c. What is the probability that in the sample, more than 62% are more likely to buy stock in a U.S. corporation, or shop at its stores, if it is making an effort to publicly talk about how it is becoming more sustainable?
- d. If a sample of 400 is taken, how does this change your answers to (a) through (c)?

7.16 According to *GMI Ratings' 2013 Women on Boards Report*, the percentage of women on U.S. boards has increased marginally in 2009–2012 and now stands at 14%, well below the values for Nordic countries and France. A number of initiatives are underway in an effort to increase the representation. For example, a network of investors, corporate leaders, and other advocates, known as the 30% coalition, is seeking to raise the proportion of female directors to that number (30%) by 2015. This study also reports that 15% of U.S. companies have three or more female board directors. (Data extracted from bit.ly/13oSFem.) If you select a random sample of 200 U.S. companies,

- a. what is the probability that the sample will have between 12% and 18% U.S. companies that have three or more female board directors?
- b. the probability is 90% that the sample percentage of U.S. companies having three or more female board directors will be

contained within what symmetrical limits of the population percentage?

- c. the probability is 95% that the sample percentage of U.S. companies having three or more female board directors will be contained within what symmetrical limits of the population percentage?

7.17 The Chartered Financial Analyst (CFA) institute reported that 51% of its U.S. members indicate that lack of ethical culture within financial firms has contributed the most to the lack of trust in the financial industry. (Source: Data extracted from *Global Market Sentiment Survey 2013*, cfa.is/YqVCKB.) Suppose that you select a sample of 100 CFA members.

- a. What is the probability that the sample percentage indicating that lack of ethical culture within financial firms has contributed the most to the lack of trust in the financial industry will be between 50% and 55%?
- b. The probability is 90% that the sample percentage will be contained within what symmetrical limits of the population percentage?
- c. The probability is 95% that the sample percentage will be contained within what symmetrical limits of the population percentage?
- d. Suppose you selected a sample of 400 CFA members. How does this change your answers in (a) through (c)?

7.18 A Pew Research Center project on the state of news media showed that the clearest pattern of news audience growth in 2012 came on digital platforms. According to Pew Research data, 39% of Americans get news online or from a mobile device in a typical day. (Data extracted from "Key Findings: State of the News Media," Pew Research Center, bit.ly/10kKUTi.)

- a. Suppose that you take a sample of 100 Americans. If the population proportion of Americans who get news online or from a mobile device in a typical day is 0.39, what is the probability that fewer than 30% in your sample will get news online or from a mobile device in a typical day?
- b. Suppose that you take a sample of 400 Americans. If the population proportion of Americans who get news online or from a mobile device in a typical day is 0.39, what is the probability that fewer than 30% in your sample will get news online or from a mobile device in a typical day?
- c. Discuss the effect of sample size on the sampling distribution of the proportion in general and the effect on the probabilities in (a) and (b).

7.4 Sampling from Finite Populations

The Central Limit Theorem and the standard errors of the mean and of the proportion are based on samples selected with replacement. However, in nearly all survey research, you sample *without* replacement from populations that are of a finite size, N . The **Section 7.4 online topic** explains how you use a **finite population correction factor** to compute the standard error of the mean and the standard error of the proportion for such samples.

USING STATISTICS

Sampling Oxford Cereals, Revisited

As the plant operations manager for Oxford Cereals, you were responsible for monitoring the amount of cereal placed in each box. To be consistent with package labeling, boxes should contain a mean of 368 grams of cereal. Thousands of boxes are produced during a shift, and weighing every single box was determined to be too time-consuming, costly, and inefficient. Instead, a sample of boxes was selected. Based on your analysis of the sample, you had to decide whether to maintain, alter, or shut down the process.

Using the concept of the sampling distribution of the mean, you were able to determine probabilities that such a sample mean could have been randomly selected from a population with a mean of 368 grams. Specifically, if a sample of

size $n = 25$ is selected from a population with a mean of 368 and standard deviation of 15,

you calculated the probability of selecting a sample with a mean of 365 grams or less to be 15.87%. If a larger sample size is selected, the sample mean should be closer to the population mean. This result was illustrated when you calculated the probability if the sample size were increased to $n = 100$. Using the larger sample size, you determined the probability of selecting a sample with a mean of 365 grams or less to be 2.28%.



Corbis

SUMMARY

You studied the sampling distribution of the sample mean and the sampling distribution of the sample proportion and their relationship to the Central Limit Theorem. You learned that the sample mean is an unbiased estimator of the popula-

tion mean, and the sample proportion is an unbiased estimator of the population proportion. In the next five chapters, the techniques of confidence intervals and tests of hypotheses commonly used for statistical inference are discussed.

REFERENCES

1. Cochran, W. G. *Sampling Techniques*, 3rd ed. New York: Wiley, 1977.
2. Microsoft Excel 2013. Redmond, WA: Microsoft Corp., 2012.
3. Minitab Release 16. State College, PA: Minitab, Inc., 2010.

KEY EQUATIONS

Population Mean

$$\mu = \frac{\sum_{i=1}^N X_i}{N} \quad (7.1)$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}} \quad (7.2)$$

Standard Error of the Mean

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (7.3)$$

Finding Z for the Sampling Distribution of the Mean

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad (7.4)$$

Finding \bar{X} for the Sampling Distribution of the Mean

$$\bar{X} = \mu + Z \frac{\sigma}{\sqrt{n}} \quad (7.5)$$

Sample Proportion

$$p = \frac{X}{n} \quad (7.6)$$

Standard Error of the Proportion

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}} \quad (7.7)$$

Finding Z for the Sampling Distribution of the Proportion

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \quad (7.8)$$

KEY TERMS

Central Limit Theorem 257

sampling distribution 251

sampling distribution of the mean 251

sampling distribution of the proportion 263

standard error of the mean 253

standard error of the proportion 263

unbiased 251

CHECKING YOUR UNDERSTANDING

7.19 Why is the sample mean an unbiased estimator of the population mean?

7.20 Why does the standard error of the mean decrease as the sample size, n , increases?

7.21 Why does the sampling distribution of the mean follow a normal distribution for a large enough sample size, even though the population may not be normally distributed?

7.22 What is the difference between a population distribution and a sampling distribution?

7.23 Under what circumstances does the sampling distribution of the proportion approximately follow the normal distribution?

CHAPTER REVIEW PROBLEMS

7.24 An industrial sewing machine uses ball bearings that are targeted to have a diameter of 0.75 inch. The lower and upper specification limits under which the ball bearing can operate are 0.74 inch (lower) and 0.76 inch (upper). Past experience has indicated that the actual diameter of the ball bearings is approximately normally distributed, with a mean of 0.753 inch and a standard deviation of 0.004 inch. If you select a random sample of 25 ball bearings, what is the probability that the sample mean is

- between the target and the population mean of 0.753?
- between the lower specification limit and the target?
- greater than the upper specification limit?
- less than the lower specification limit?
- The probability is 93% that the sample mean diameter will be greater than what value?

7.25 The fill amount of bottles of a soft drink is normally distributed, with a mean of 2.0 liters and a standard deviation of 0.05 liter. If you select a random sample of 25 bottles, what is the probability that the sample mean will be

- between 1.99 and 2.0 liters?
- below 1.98 liters?
- greater than 2.01 liters?
- The probability is 99% that the sample mean amount of soft drink will be at least how much?
- The probability is 99% that the sample mean amount of soft drink will be between which two values (symmetrically distributed around the mean)?

7.26 An orange juice producer buys oranges from a large orange grove that has one variety of orange. The amount of juice squeezed from these oranges is approximately normally distributed, with a mean of 4.70 ounces and a standard deviation of 0.40 ounce. Suppose that you select a sample of 25 oranges.

- What is the probability that the sample mean amount of juice will be at least 4.60 ounces?
- The probability is 70% that the sample mean amount of juice will be contained between what two values symmetrically distributed around the population mean?
- The probability is 77% that the sample mean amount of juice will be greater than what value?

7.27 In Problem 7.26, suppose that the mean amount of juice squeezed is 5.0 ounces.

- What is the probability that the sample mean amount of juice will be at least 4.60 ounces?
- The probability is 70% that the sample mean amount of juice will be contained between what two values symmetrically distributed around the population mean?
- The probability is 77% that the sample mean amount of juice will be greater than what value?

7.28 The stock market in Mexico reported strong returns in 2012. The population of stocks earned a mean return of 17.87% in 2012. (Data extracted from asxiq.com/blog/stock-markets-in-the-world-returns-in-2012/.) Assume that the returns for stocks on the Mexican stock market were distributed as a normal random

variable, with a mean of 17.87 and a standard deviation of 20. If you selected a random sample of 16 stocks from this population, what is the probability that the sample would have a mean return

- less than 0 (i.e., a loss)?
- between -10 and 10?
- greater than 10?

7.29 The article mentioned in Problem 7.28 reported that the stock market in China had a mean return of 1.54% in 2012. Assume that the returns for stocks on the Chinese stock market were distributed as a normal random variable, with a mean of 1.54 and a standard deviation of 10. If you select an individual stock from this population, what is the probability that it would have a return

- less than 0 (i.e., a loss)?
- between -10 and -20?
- greater than -5?

If you selected a random sample of four stocks from this population, what is the probability that the sample would have a mean return

- less than 0 (a loss)?
- between -10 and -20?
- greater than -5?
- Compare your results in parts (d) through (f) to those in (a) through (c).

7.30 (Class Project) The table of random numbers is an example of a uniform distribution because each digit is equally likely to occur. Starting in the row corresponding to the day of the month in which you were born, use a table of random numbers (Table E.1) to take one digit at a time.

Select five different samples each of $n = 2$, $n = 5$, and $n = 10$. Compute the sample mean of each sample. Develop a frequency distribution of the sample means for the results of the entire class, based on samples of sizes $n = 2$, $n = 5$, and $n = 10$.

What can be said about the shape of the sampling distribution for each of these sample sizes?

7.31 (Class Project) Toss a coin 10 times and record the number of heads. If each student performs this experiment five times, a frequency distribution of the number of heads can be developed from the results of the entire class. Does this distribution seem to approximate the normal distribution?

7.32 (Class Project) The number of cars waiting in line at a car wash is distributed as follows:

Number of Cars	Probability
0	0.25
1	0.40
2	0.20
3	0.10
4	0.04
5	0.01

You can use a table of random numbers (Table E.1) to select samples from this distribution by assigning numbers as follows:

- Start in the row corresponding to the day of the month in which you were born.
- Select a two-digit random number.
- If you select a random number from 00 to 24, record a length of 0; if from 25 to 64, record a length of 1; if from 65 to 84, record a length of 2; if from 85 to 94, record a length of 3; if from 95 to 98, record a length of 4; if 99, record a length of 5.

Select samples of $n = 2$, $n = 5$, and $n = 10$. Compute the mean for each sample. For example, if a sample of size 2 results in the random numbers 18 and 46, these would correspond to lengths 0 and 1, respectively, producing a sample mean of 0.5. If each student selects five different samples for each sample size, a frequency distribution of the sample means (for each sample size) can be developed from the results of the entire class. What conclusions can you reach concerning the sampling distribution of the mean as the sample size is increased?

7.33 (Class Project) Using a table of random numbers (Table E.1), simulate the selection of different-colored balls from a bowl, as follows:

- Start in the row corresponding to the day of the month in which you were born.
- Select one-digit numbers.
- If a random digit between 0 and 6 is selected, consider the ball white; if a random digit is a 7, 8, or 9, consider the ball red.

Select samples of $n = 10$, $n = 25$, and $n = 50$ digits. In each sample, count the number of white balls and compute the proportion of white balls in the sample. If each student in the class selects five different samples for each sample size, a frequency distribution of the proportion of white balls (for each sample size) can be developed from the results of the entire class. What conclusions can you reach about the sampling distribution of the proportion as the sample size is increased?

7.34 (Class Project) Suppose that step 3 of Problem 7.33 uses the following rule: "If a random digit between 0 and 8 is selected, consider the ball to be white; if a random digit of 9 is selected, consider the ball to be red." Compare and contrast the results in this problem and those in Problem 7.33.

CASES FOR CHAPTER 7

Managing Ashland MultiComm Services

Continuing the quality improvement effort first described in the Chapter 6 Managing Ashland MultiComm Services case, the target upload speed for AMS Internet service subscribers has been monitored. As before, upload speeds are measured on a standard scale in which the target value is 1.0. Data collected over the past year indicate that the upload speeds are approximately normally distributed, with a mean of 1.005 and a standard deviation of 0.10.

1. Each day, at 25 random times, the upload speed is measured. Assuming that the distribution has not changed from what it was in the past year, what is the probability that the mean upload speed is

- a. less than 1.0?
- b. between 0.95 and 1.0?
- c. between 1.0 and 1.05?
- d. less than 0.95 or greater than 1.05?
- e. Suppose that the mean upload speed of today's sample of 25 is 0.952. What conclusion can you reach about the upload speed today based on this result? Explain.

2. Compare the results of AMS Problem 1 (a) through (d) to those of AMS Problem 1 in Chapter 6 on page 246. What conclusions can you reach concerning the differences?

Digital Case

Apply your knowledge about sampling distributions in this Digital Case, which reconsiders the Oxford Cereals Using Statistics scenario.

The advocacy group Consumers Concerned About Cereal Cheaters (CCACC) suspects that cereal companies, including Oxford Cereals, are cheating consumers by packaging cereals at less than labeled weights. Recently, the group investigated the package weights of two popular Oxford brand cereals. Open [CCACC.pdf](#) to examine the group's claims and supporting data, and then answer the following questions:

1. Are the data collection procedures that the CCACC uses to form its conclusions flawed? What procedures could the group follow to make its analysis more rigorous?
2. Assume that the two samples of five cereal boxes (one sample for each of two cereal varieties) listed on the CCACC website were collected randomly by organization members. For each sample,

- a. calculate the sample mean.
- b. assuming that the standard deviation of the process is 15 grams and the population mean is 368 grams, calculate the percentage of all samples for each process that have a sample mean less than the value you calculated in (a).
- c. assuming that the standard deviation is 15 grams, calculate the percentage of individual boxes of cereal that have a weight less than the value you calculated in (a).
- 3. What, if any, conclusions can you form by using your calculations about the filling processes for the two different cereals?
- 4. A representative from Oxford Cereals has asked that the CCACC take down its page discussing shortages in Oxford Cereals boxes. Is this request reasonable? Why or why not?
- 5. Can the techniques discussed in this chapter be used to prove cheating in the manner alleged by the CCACC? Why or why not?

CHAPTER 7 EXCEL GUIDE

EG7.1 SAMPLING DISTRIBUTIONS

There are no Excel Guide instructions for this section.

EG7.2 SAMPLING DISTRIBUTION of the MEAN

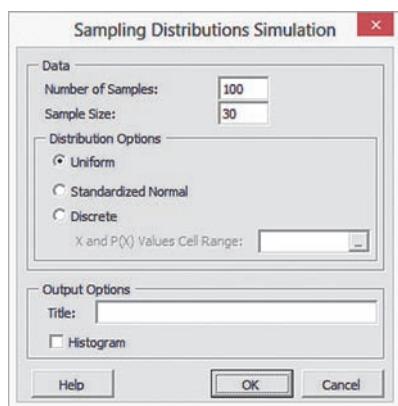
Key Technique Use an add-in procedure to create a simulated sampling distribution and use the **RAND()** function to create lists of random numbers.

Example Create a simulated sampling distribution that consists of 100 samples of $n = 30$ from a uniformly distributed population.

PHStat Use Sampling Distributions Simulation.

For the example, select **PHStat → Sampling → Sampling Distributions Simulation**. In the procedure's dialog box (shown below):

1. Enter **100** as the **Number of Samples**.
2. Enter **30** as the **Sample Size**.
3. Click **Uniform**.
4. Enter a **Title** and click **OK**.



The procedure inserts a new worksheet in which the sample means, overall mean, and standard error of the mean can be found starting in row 34.

In-Depth Excel Use the **SDS worksheet** of the **SDS workbook** as a model.

For the example, in a new worksheet, first enter a title in cell A1. Then enter the formula **=RAND()** in cell A2 and then copy the formula down 30 rows and across 100 columns (through

column CV). Then select this cell range (**A2:CV31**) and use **copy and paste values** as discussed in Appendix Section B.4.

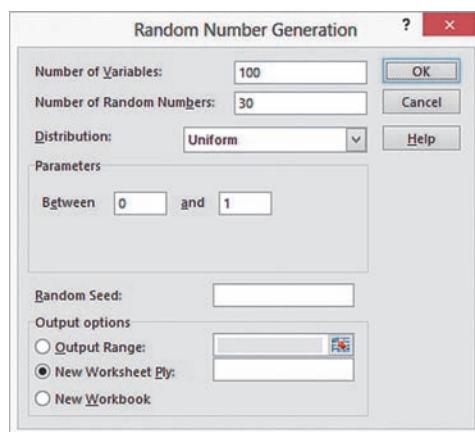
Use the formulas that appear in rows 33 through 37 in the **SDS_FORMULAS worksheet** of the **SDS workbook** as models if you want to compute sample means, the overall mean, and the standard error of the mean.

Analysis ToolPak Use Random Number Generation.

For the example, select **Data → Data Analysis**. In the Data Analysis dialog box, select **Random Number Generation** from the **Analysis Tools** list and then click **OK**.

In the procedure's dialog box (shown below):

1. Enter **100** as the **Number of Variables**.
2. Enter **30** as the **Number of Random Numbers**.
3. Select **Uniform** from the **Distribution** drop-down list.
4. Keep the **Parameters** values as is.
5. Click **New Worksheet Ply** and then click **OK**.



If, for other problems, you select **Discrete** in step 3, you must be open to a worksheet that contains a cell range of X and $P(X)$ values. Enter this cell range as the **Value and Probability Input Range** (not shown when **Uniform** has been selected) in the **Parameters** section of the dialog box.

Use the formulas that appear in rows 33 through 37 in the **SDS_FORMULAS worksheet** of the **SDS workbook** as models if you want to compute sample means, the overall mean, and the standard error of the mean.

EG7.3 SAMPLING DISTRIBUTION of the PROPORTION

There are no Excel Guide instructions for this section.

CHAPTER 7 MINITAB GUIDE

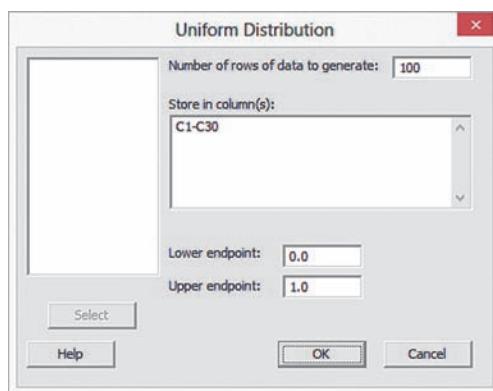
MG7.1 SAMPLING DISTRIBUTIONS

There are no Minitab Guide instructions for this section.

MG7.2 SAMPLING DISTRIBUTION of the MEAN

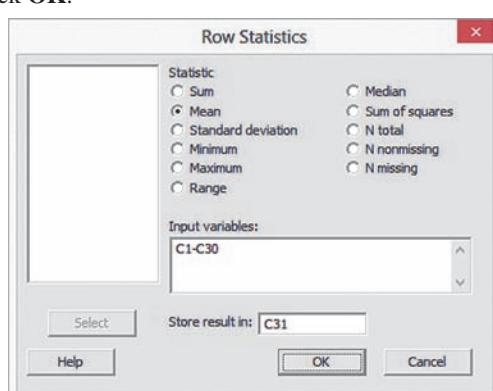
Use **Uniform** to create a simulated sampling distribution from a uniformly distributed population. For example, to create 100 samples of $n = 30$ from a uniformly distributed population, open to a new worksheet. Select **Calc → Random Data → Uniform**. In the Uniform Distribution dialog box (shown below):

1. Enter **100** in the **Number of rows of data to generate** box.
2. Enter **C1-C30** in the **Store in column(s)** box (to store the results in the first 30 columns).
3. Enter **0.0** in the **Lower endpoint** box.
4. Enter **1.0** in the **Upper endpoint** box.
5. Click **OK**.



The 100 samples of $n = 30$ are entered *row-wise* in columns C1 through C30, an exception to the rule used in this book to enter data column-wise. (Row-wise data facilitates the computation of means.) While still opened to the worksheet with the 100 samples, enter **Sample Means** as the name of column **C31**. Select **Calc → Row Statistics**. In the Row Statistics dialog box (shown below):

6. Click **Mean**.
7. Enter **C1-C30** in the **Input variables** box.
8. Enter **C31** in the **Store result in** box.
9. Click **OK**.



10. With the mean for each of the 100 row-wise samples in column C31, select **Stat → Basic Statistics → Display Descriptive Statistics**.

11. In the Display Descriptive Statistics dialog box, enter **C31** in the **Variables** box and click **Statistics**.
12. In the Display Descriptive Statistics - Statistics dialog box, select **Mean** and **Standard deviation** and then click **OK**.
13. Back in the Display Descriptive Statistics dialog box, click **OK**.

While still open to the worksheet created in steps 1 through 13, select **Graph → Histogram** and in the Histograms dialog box, click **Simple** and then click **OK**. In the Histogram - Simple dialog box:

1. Enter **C31** in the **Graph variables** box.
2. Click **OK**.

Sampling from Normally Distributed Populations

Use **Normal** to create a simulated sampling distribution from a normally distributed population. For example, to create 100 samples of $n = 30$ from a normally distributed population, open to a new worksheet. Select **Calc → Random Data → Normal**. In the Normal Distribution dialog box:

1. Enter **100** in the **Number of rows of data to generate** box.
2. Enter **C1-C30** in the **Store in column(s)** box (to store the results in the first 30 columns).
3. Enter a value for μ in the **Mean** box.
4. Enter a value for σ in the **Standard deviation** box.
5. Click **OK**.

The 100 samples of $n = 30$ are entered row-wise in columns C1 through C30. To compute statistics, select **Calc → Row Statistics** and follow steps 6 through 13 from the set of instructions for a uniformly distributed population.

MG7.3 SAMPLING DISTRIBUTION of the PROPORTION

There are no Minitab Guide instructions for this section.

CHAPTER

8

Confidence Interval Estimation

CONTENTS

- 8.1 Confidence Interval Estimate for the Mean (σ Known)
- 8.2 Confidence Interval Estimate for the Mean (σ Unknown)
- 8.3 Confidence Interval Estimate for the Proportion
- 8.4 Determining Sample Size
- 8.5 Confidence Interval Estimation and Ethical Issues
- 8.6 Application of Confidence Interval Estimation in Auditing (*online*)
- 8.7 Estimation and Sample Size Estimation for Finite Populations (*online*)
- 8.8 Bootstrapping (*online*)

USING STATISTICS: Getting Estimates at Ricknel Home Centers, Revisited

CHAPTER 8 EXCEL GUIDE

CHAPTER 8 MINITAB GUIDE

OBJECTIVES

- To construct and interpret confidence interval estimates for the mean and the proportion
- To determine the sample size necessary to develop a confidence interval estimate for the mean or proportion

USING STATISTICS

Getting Estimates at Ricknel Home Centers

As a member of the AIS team at Ricknel Home Centers (see page 185), you have already examined the probability of discovering questionable, or “tagged,” invoices. Now you have been assigned the task of auditing the accuracy of the integrated inventory management and point of sale component of the firm’s retail management system.

You could review the contents of each and every inventory and transactional record to check the accuracy of this system, but such a detailed review would be time-consuming and costly. Could you use statistical inference techniques to reach conclusions about the population of all records from a relatively small sample collected during an audit? At the end of each month, you could select a sample of the sales invoices to estimate population parameters such as

- The mean dollar amount listed on the sales invoices for the month
- The proportion of invoices that contain errors that violate the internal control policy of the warehouse

If you used the sampling technique, how accurate would the results from the sample be? How would you use the results you generate? How could you be certain that the sample size is large enough to give you the information you need?



Mangostock/Shutterstock

In Section 7.2, you used the Central Limit Theorem and knowledge of the population distribution to determine the percentage of sample means that are within certain distances of the population mean. For instance, in the cereal-filling example used throughout Chapter 7 (see Example 7.4 on page 257), you can conclude that 95% of all sample means are between 362.12 and 373.88 grams. This is an example of *deductive* reasoning because the conclusion is based on taking something that is true in general (for the population) and applying it to something specific (the sample means).

Getting the results that Ricknel Home Centers needs requires *inductive* reasoning. Inductive reasoning lets you use some specifics to make broader generalizations. You cannot guarantee that the broader generalizations are absolutely correct, but with a careful choice of the specifics and a rigorous methodology, you can get useful conclusions. As a Ricknel accountant, you need to use inferential statistics, which uses sample results (the “some specifics”) to *estimate* (the making of “broader generalizations”) unknown population parameters such as a population mean or a population proportion. Note that statisticians use the word *estimate* in the same sense of the everyday usage: something you are reasonably certain about but cannot flatly say is absolutely correct.

You estimate population parameters by using either point estimates or interval estimates. A **point estimate** is the value of a single sample statistic, such as a sample mean. A **confidence interval estimate** is a range of numbers, called an *interval*, constructed around the point estimate. The confidence interval is constructed such that the probability that the interval includes the population parameter is known.

Suppose you want to estimate the mean GPA of all the students at your university. The mean GPA for all the students is an unknown population mean, denoted by μ . You select a sample of students and compute the sample mean, denoted by \bar{X} , to be 2.80. As a *point estimate* of the population mean, μ , you ask how accurate is the 2.80 value as an estimate of the population mean, μ ? By taking into account the variability from sample to sample (see Section 7.2, concerning the sampling distribution of the mean), you can construct a confidence interval estimate for the population mean to answer this question.

When you construct a confidence interval estimate, you indicate the confidence of correctly estimating the value of the population parameter, μ . This allows you to say that there is a specified confidence that μ is somewhere in the range of numbers defined by the interval.

After studying this chapter, you might find that a 95% confidence interval for the mean GPA at your university is $2.75 \leq \mu \leq 2.85$. You can interpret this interval estimate by stating that you are 95% confident that the mean GPA at your university is between 2.75 and 2.85.

In this chapter, you learn to construct a confidence interval for both the population mean and population proportion. You also learn how to determine the sample size that is necessary to construct a confidence interval of a desired width.

8.1 Confidence Interval Estimate for the Mean (σ Known)

In Section 7.2, you used the Central Limit Theorem and knowledge of the population distribution to determine the percentage of sample means that are within certain distances of the population mean. Suppose that in the cereal-filling example you wished to estimate the population mean, using the information from a single sample. Thus, rather than taking $\mu \pm (1.96)(\sigma/\sqrt{n})$ to find the upper and lower limits around μ , as in Section 7.2, you substitute the sample mean, \bar{X} , for the unknown μ and use $\bar{X} \pm (1.96)(\sigma/\sqrt{n})$ as an interval to estimate the unknown μ . Although in practice you select a single sample of n values and compute the mean, \bar{X} , in order to understand the full meaning of the interval estimate, you need to examine a hypothetical set of all possible samples of n values.

Suppose that a sample of $n = 25$ cereal boxes has a mean of 362.3 grams and a standard deviation of 15 grams. The interval developed to estimate μ is $362.3 \pm (1.96)(15)/(\sqrt{25})$, or 362.3 ± 5.88 . The estimate of μ is

$$356.42 \leq \mu \leq 368.18$$

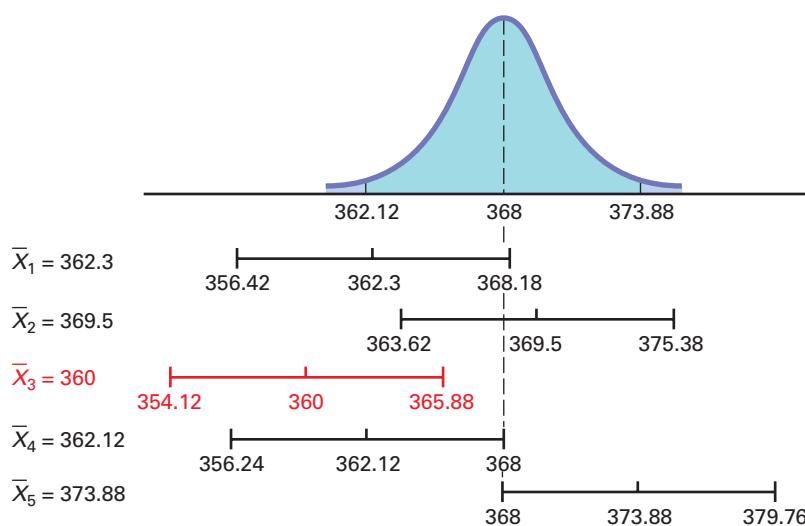
Student Tip

Remember, the confidence interval is for the population mean not the sample mean.

Because the population mean, μ (equal to 368), is included within the interval, this sample results in a correct statement about μ (see Figure 8.1).

FIGURE 8.1

Confidence interval estimates for five different samples of $n = 25$ taken from a population where $\mu = 368$ and $\sigma = 15$



To continue this hypothetical example, suppose that for a different sample of $n = 25$ boxes, the mean is 369.5. The interval developed from this sample is

$$369.5 \pm (1.96)(15)/(\sqrt{25})$$

or 369.5 ± 5.88 . The estimate is

$$363.62 \leq \mu \leq 375.38$$

Because the population mean, μ (equal to 368), is also included within this interval, this statement about μ is correct.

Now, before you begin to think that correct statements about μ are always made by developing a confidence interval estimate, suppose a third hypothetical sample of $n = 25$ boxes is selected and the sample mean is equal to 360 grams. The interval developed here is $360 \pm (1.96)(15)/(\sqrt{25})$, or 360 ± 5.88 . In this case, the estimate of μ is

$$354.12 \leq \mu \leq 365.88$$

This estimate is *not* a correct statement because the population mean, μ , is not included in the interval developed from this sample (see Figure 8.1). Thus, for some samples, the interval estimate for μ is correct, but for others it is incorrect. In practice, only one sample is selected, and because the population mean is unknown, you cannot determine whether the interval estimate is correct. To resolve this, you need to determine the proportion of samples producing intervals that result in correct statements about the population mean, μ . To do this, consider two other hypothetical samples: the case in which $\bar{X} = 362.12$ grams and the case in which $\bar{X} = 373.88$ grams. If $\bar{X} = 362.12$, the interval is $362.12 \pm (1.96)(15)/(\sqrt{25})$, or 362.12 ± 5.88 . This leads to the following interval:

$$356.24 \leq \mu \leq 368.00$$

Because the population mean of 368 is at the upper limit of the interval, the statement is correct (see Figure 8.1).

When $\bar{X} = 373.88$, the interval is $373.88 \pm (1.96)(15)/(\sqrt{25})$, or 373.88 ± 5.88 . The interval estimate for the mean is

$$368.00 \leq \mu \leq 379.76$$

In this case, because the population mean of 368 is included at the lower limit of the interval, the statement is correct.

In Figure 8.1, you see that when the sample mean falls somewhere between 362.12 and 373.88 grams, the population mean is included *somewhere* within the interval. In Example 7.4 on page 257, you found that 95% of the sample means are between 362.12 and 373.88 grams. Therefore, 95% of all samples of $n = 25$ boxes have sample means that will result in intervals that include the population mean.

Because, in practice, you select only one sample of size n , and μ is unknown, you never know for sure whether your specific interval includes the population mean. However, if you take all possible samples of n and compute their 95% confidence intervals, 95% of the intervals will include the population mean, and only 5% of them will not. In other words, you have 95% confidence that the population mean is somewhere in your interval.

Consider once again the first sample discussed in this section. A sample of $n = 25$ boxes had a sample mean of 362.3 grams. The interval constructed to estimate μ is

$$362.3 \pm (1.96)(15)/(\sqrt{25})$$

$$362.3 \pm 5.88$$

$$356.42 \leq \mu \leq 368.18$$

The interval from 356.42 to 368.18 is referred to as a *95% confidence interval*. The following contains an interpretation of the interval that most business professionals will understand. (For a technical discussion of different ways to interpret confidence intervals, see reference 4.)

“I am 95% confident that the mean amount of cereal in the population of boxes is somewhere between 356.42 and 368.18 grams.”

To help you understand the meaning of the confidence interval, consider the order-filling process at a website. Filling orders consists of several steps, including receiving an order, picking the parts of the order, checking the order, packing, and shipping the order. The file **Order** contains the time, in minutes, to fill orders for a population of $N = 200$ orders on a recent day. Although in practice the population characteristics are rarely known, for this population of orders, the mean, μ , is known to be equal to 69.637 minutes; the standard deviation, σ , is known to be equal to 10.411 minutes; and the population is normally distributed. To illustrate how the sample mean and sample standard deviation can vary from one sample to another, 20 different samples of $n = 10$ were selected from the population of 200 orders, and the sample mean and sample standard deviation (and other statistics) were calculated for each sample. Figure 8.2 shows these results.

FIGURE 8.2

Sample statistics and 95% confidence intervals for 20 samples of $n = 10$ randomly selected from the population of $N = 200$ orders

Sample	n	Mean	Std Dev	Minimum	Median	Maximum	Range	95% Conf. Int.
S01	10	74.15	13.39	56.10	76.85	97.70	41.60	(67.70, 80.60)
S02	10	61.10	10.60	46.80	61.35	79.50	32.70	(54.65, 67.55)
S03	10	74.36	6.50	62.50	74.50	84.00	21.50	(67.91, 80.81)
S04	10	70.40	12.80	47.20	70.95	84.00	36.80	(63.95, 76.85)
S05	10	62.18	10.85	47.10	59.70	84.00	36.90	(55.73, 68.63)
S06	10	67.03	9.68	51.10	69.60	83.30	32.20	(60.58, 73.48)
S07	10	69.03	8.81	56.60	68.85	83.70	27.10	(62.58, 75.48)
S08	10	72.30	11.52	54.20	71.35	87.00	32.80	(65.85, 78.75)
S09	10	68.18	14.10	50.10	69.95	86.20	36.10	(61.73, 74.63)
S10	10	66.67	9.08	57.10	64.65	86.10	29.00	(60.22, 73.12)
S11	10	72.42	9.76	59.60	74.65	86.10	26.50	(65.97, 78.87)
S12	10	76.26	11.69	50.10	80.60	87.00	36.90	(69.81, 82.71)
S13	10	65.74	12.11	47.10	62.15	86.10	39.00	(59.29, 72.19)
S14	10	69.99	10.97	51.00	73.40	84.60	33.60	(63.54, 76.44)
S15	10	75.76	8.60	61.10	75.05	87.80	26.70	(69.31, 82.21)
S16	10	67.94	9.19	56.70	67.70	87.80	31.10	(61.49, 74.39)
S17	10	71.05	10.48	50.10	71.15	86.20	36.10	(64.60, 77.50)
S18	10	71.68	7.96	55.60	72.35	82.60	27.00	(65.23, 78.13)
S19	10	70.97	9.83	54.40	70.05	84.00	30.20	(64.52, 77.42)
S20	10	74.48	8.80	62.00	76.25	85.70	23.70	(68.03, 80.93)

From Figure 8.2, you can see the following:

- The sample statistics differ from sample to sample. The sample means vary from 61.10 to 76.26 minutes, the sample standard deviations vary from 6.50 to 14.10 minutes, the sample medians vary from 59.70 to 80.60 minutes, and the sample ranges vary from 21.50 to 41.60 minutes.
- Some of the sample means are greater than the population mean of 69.637 minutes, and some of the sample means are less than the population mean.
- Some of the sample standard deviations are greater than the population standard deviation of 10.411 minutes, and some of the sample standard deviations are less than the population standard deviation.
- The variation in the sample ranges is much more than the variation in the sample standard deviations.

The variation of sample statistics from sample to sample is called *sampling error*. **Sampling error** is the variation that occurs due to selecting a single sample from the population. The size of the sampling error is primarily based on the amount of variation in the population and on the sample size. Large samples have less sampling error than small samples, but large samples cost more to select.

The last column of Figure 8.2 contains 95% confidence interval estimates of the population mean order-filling time, based on the results of those 20 samples of $n = 10$. Begin by examining the first sample selected. The sample mean is 74.15 minutes, and the interval estimate for the population mean is 67.70 to 80.60 minutes. In a typical study, you would not know for sure whether this interval estimate is correct because you rarely know the value of the population mean. However, for this example *concerning the order-filling times*, the population mean is known to be 69.637 minutes. If you examine the interval 67.70 to 80.60 minutes, you see that the population mean of 69.637 minutes is located *between* these lower and upper limits. Thus, the first sample provides a correct estimate of the population mean in the form of an interval estimate. Looking over the other 19 samples, you see that similar results occur for all the other samples *except* for samples 2, 5, and 12. For each of the intervals generated (other than samples 2, 5, and 12), the population mean of 69.637 minutes is located *somewhere* within the interval.

For sample 2, the sample mean is 61.10 minutes, and the interval is 54.65 to 67.55 minutes; for sample 5, the sample mean is 62.18, and the interval is between 55.73 and 68.63; for sample 12, the sample mean is 76.26, and the interval is between 69.81 and 82.71 minutes. The population mean of 69.637 minutes is *not* located within any of these intervals, and the estimate of the population mean made using these intervals is incorrect. Although 3 of the 20 intervals did not include the population mean, if you had selected all the possible samples of $n = 10$ from a population of $N = 200$, 95% of the intervals would include the population mean.

In some situations, you might want a higher degree of confidence of including the population mean within the interval (such as 99%). In other cases, you might accept less confidence (such as 90%) of correctly estimating the population mean. In general, the **level of confidence** is symbolized by $(1 - \alpha) \times 100\%$, where α is the proportion in the tails of the distribution that is outside the confidence interval. The proportion in the upper tail of the distribution is $\alpha/2$, and the proportion in the lower tail of the distribution is $\alpha/2$. You use Equation (8.1) to construct a $(1 - \alpha) \times 100\%$ confidence interval estimate for the mean with σ known.

CONFIDENCE INTERVAL FOR THE MEAN (σ KNOWN)

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

or

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (8.1)$$

where $Z_{\alpha/2}$ is the value corresponding to an upper-tail probability of $\alpha/2$ from the standardized normal distribution (i.e., a cumulative area of $1 - \alpha/2$).

The value of $Z_{\alpha/2}$ needed for constructing a confidence interval is called the **critical value** for the distribution. 95% confidence corresponds to an α value of 0.05. The critical Z value corresponding to a cumulative area of 0.975 is 1.96 because there is 0.025 in the upper tail of the distribution, and the cumulative area less than $Z = 1.96$ is 0.975.

There is a different critical value for each level of confidence, $1 - \alpha$. A level of confidence of 95% leads to a Z value of 1.96 (see Figure 8.3). 99% confidence corresponds to an α value of 0.01. The Z value is approximately 2.58 because the upper-tail area is 0.005 and the cumulative area less than $Z = 2.58$ is 0.995 (see Figure 8.4).

FIGURE 8.3

Normal curve for determining the Z value needed for 95% confidence

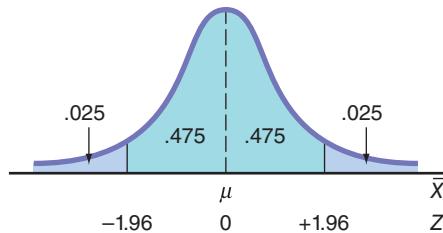
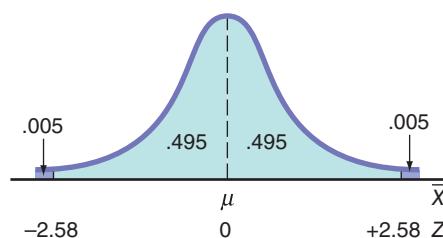


FIGURE 8.4

Normal curve for determining the Z value needed for 99% confidence



Now that various levels of confidence have been considered, why not make the confidence level as close to 100% as possible? Before doing so, you need to realize that any increase in the level of confidence is achieved only by widening (and making less precise) the confidence interval. There is no “free lunch” here. You would have more confidence that the population mean is within a broader range of values; however, this might make the interpretation of the confidence interval less useful. The trade-off between the width of the confidence interval and the level of confidence is discussed in greater depth in the context of determining the sample size in Section 8.4. Example 8.1 illustrates the application of the confidence interval estimate.

EXAMPLE 8.1

Estimating the Mean Paper Length with 95% Confidence

A paper manufacturer has a production process that operates continuously throughout an entire production shift. The paper is expected to have a mean length of 11 inches, and the standard deviation of the length is 0.02 inch. At periodic intervals, a sample is selected to determine whether the mean paper length is still equal to 11 inches or whether something has gone wrong in the production process to change the length of the paper produced. You select a random sample of 100 sheets, and the mean paper length is 10.998 inches. Construct a 95% confidence interval estimate for the population mean paper length.

SOLUTION Using Equation (8.1) on page 276, with $Z_{\alpha/2} = 1.96$ for 95% confidence,

$$\begin{aligned}\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &= 10.998 \pm (1.96) \frac{0.02}{\sqrt{100}} \\ &= 10.998 \pm 0.0039 \\ 10.9941 &\leq \mu \leq 11.0019\end{aligned}$$

Thus, with 95% confidence, you conclude that the population mean is between 10.9941 and 11.0019 inches. Because the interval includes 11, the value indicating that the production process is working properly, you have no reason to believe that anything is wrong with the production process.

Example 8.2 illustrates the effect of using a 99% confidence interval.

EXAMPLE 8.2

Estimating the Mean Paper Length with 99% Confidence

Construct a 99% confidence interval estimate for the population mean paper length.

SOLUTION Using Equation (8.1) on page 276, with $Z_{\alpha/2} = 2.58$ for 99% confidence,

$$\begin{aligned}\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &= 10.998 \pm (2.58) \frac{0.02}{\sqrt{100}} \\ &= 10.998 \pm 0.00516 \\ 10.9928 &\leq \mu \leq 11.0032\end{aligned}$$

Once again, because 11 is included within this wider interval, you have no reason to believe that anything is wrong with the production process.

As discussed in Section 7.2, the sampling distribution of the sample mean, \bar{X} , is normally distributed if the population for your characteristic of interest, X , follows a normal distribution. And if the population of X does not follow a normal distribution, the Central Limit Theorem almost always ensures that \bar{X} is approximately normally distributed when n is large. However, when dealing with a small sample size and a population that does not follow a normal distribution, the sampling distribution of \bar{X} is not normally distributed, and therefore the confidence interval discussed in this section is inappropriate. In practice, however, as long as the sample size is large enough and the population is not very skewed, you can use the confidence interval defined in Equation (8.1) to estimate the population mean when σ is known. To assess the assumption of normality, you can evaluate the shape of the sample data by constructing a histogram, stem-and-leaf display, boxplot, or normal probability plot.

Student Tip

Because understanding the confidence interval concept is very important when reading the rest of this book, review this section carefully to understand the underlying concept—even if you never have a practical reason to use the confidence interval estimate of the mean (σ known) method.

Can You Ever Know the Population Standard Deviation?

To solve Equation (8.1), you must know the value for σ , the population standard deviation. To know σ implies that you know all the values in the entire population. (How else would you know the value of this population parameter?) If you knew all the values in the entire population, you could directly compute the population mean. There would be no need to use the *inductive* reasoning of inferential statistics to *estimate* the population mean. In other words, if you know σ , you really do not have a need to use Equation (8.1) to construct a confidence interval estimate of the mean (σ known).

More significantly, in virtually all real-world business situations, you would never know the standard deviation of the population. In business situations, populations are often too large to examine all the values. So why study the confidence interval estimate of the mean (σ known) at all? This method serves as an important introduction to the concept of a confidence interval because it uses the normal distribution, which has already been thoroughly discussed in Chapters 6 and 7. In the next section, you will see that constructing a confidence interval estimate when σ is not known requires another distribution (the t distribution) not previously mentioned in this book.

Problems for Section 8.1

LEARNING THE BASICS

8.1 If $\bar{X} = 85$, $\sigma = 8$, and $n = 64$, construct a 95% confidence interval estimate for the population mean, μ .

8.2 If $\bar{X} = 125$, $\sigma = 24$, and $n = 36$, construct a 99% confidence interval estimate for the population mean, μ .

8.3 Why is it not possible in Example 8.1 on page 277 to have 100% confidence? Explain.

8.4 Is it true in Example 8.1 on page 277 that you do not know for sure whether the population mean is between 10.9941 and 11.0019 inches? Explain.

APPLYING THE CONCEPTS

8.5 A market researcher selects a simple random sample of $n = 100$ Twitter users from a population of over 100 million Twitter registered users. After analyzing the sample, she states that she has 95% confidence that the mean time spent on the site per day is between 15 and 57 minutes. Explain the meaning of this statement.

8.6 Suppose that you are going to collect a set of data, either from an entire population or from a random sample taken from that population.

- Which statistical measure would you compute first: the mean or the standard deviation? Explain.
- What does your answer to (a) tell you about the “practicality” of using the confidence interval estimate formula given in Equation (8.1)?

8.7 Consider the confidence interval estimate discussed in Problem 8.5. Suppose the population mean time spent on the site is 36 minutes a day. Is the confidence interval estimate stated in Problem 8.5 correct? Explain.

8.8 You are working as an assistant to the dean of institutional research at your university. The dean wants to survey members of the alumni association who obtained their baccalaureate degrees five years ago to learn what their starting salaries were in their first full-time job after receiving their degrees. A sample of 100 alumni is to be randomly selected from the list of 2,500 graduates in that class. If the dean’s goal is to construct a 95% confidence interval estimate for the population mean starting salary, why is it not possible that you will be able to use Equation (8.1) on page 276 for this purpose? Explain.

8.9 A bottled water distributor wants to estimate the amount of water contained in 1-gallon bottles purchased from a nationally known water bottling company. The water bottling company’s specifications state that the standard deviation of the amount of water is equal to 0.02 gallon. A random sample of 50 bottles is selected, and the sample mean amount of water per 1-gallon bottle is 0.995 gallon.

- Construct a 99% confidence interval estimate for the population mean amount of water included in a 1-gallon bottle.
- On the basis of these results, do you think that the distributor has a right to complain to the water bottling company? Why?
- Must you assume that the population amount of water per bottle is normally distributed here? Explain.
- Construct a 95% confidence interval estimate. How does this change your answer to (b)?



8.10 The operations manager at a compact fluorescent light bulb (CFL) factory needs to estimate the mean life of a large shipment of CFLs. The manufacturer’s specifications are that the standard deviation is 1,000 hours. A random sample of 64 CFLs indicated a sample mean life of 7,500 hours.

- Construct a 95% confidence interval estimate for the population mean life of compact fluorescent light bulbs in this shipment.
- Do you think that the manufacturer has the right to state that the compact fluorescent light bulbs have a mean life of 8,000 hours? Explain.
- Must you assume that the population compact fluorescent light bulb life is normally distributed? Explain.
- Suppose that the standard deviation changes to 800 hours. What are your answers in (a) and (b)?

8.2 Confidence Interval Estimate for the Mean (σ Unknown)

In the previous section, you learned that in most business situations, you do not know σ , the population standard deviation. This section discusses a method of constructing a confidence interval estimate of μ that uses the sample statistic S as an estimate of the population parameter σ .

Student's t Distribution

At the start of the twentieth century, William S. Gosset was working at Guinness in Ireland, trying to help brew better beer less expensively (see reference 5). As he had only small samples to study, he needed to find a way to make inferences about means without having to know σ . Writing under the pen name “Student,”¹ Gosset solved this problem by developing what today is known as the **Student's t distribution**, or the t distribution.

If the random variable X is normally distributed, then the following statistic:

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

has a t distribution with $n - 1$ **degrees of freedom**. This expression has the same form as the Z statistic in Equation (7.4) on page 255, except that S is used to estimate the unknown σ .

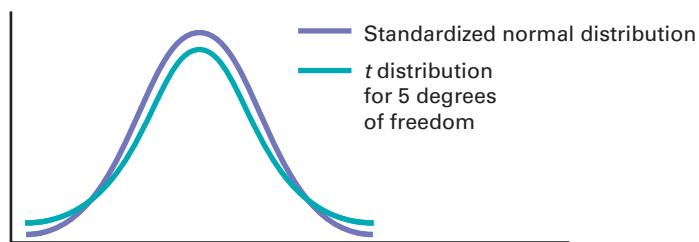
¹Guinness considered all research conducted to be proprietary and a trade secret. The firm prohibited its employees from publishing their results. Gosset circumvented this ban by using the pen name “Student” to publish his findings.

Properties of the *t* Distribution

The *t* distribution is very similar in appearance to the standardized normal distribution. Both distributions are symmetrical and bell-shaped, with the mean and the median equal to zero. However, because S is used to estimate the unknown σ , the values of *t* are more variable than those for *Z*. Therefore, the *t* distribution has more area in the tails and less in the center than does the standardized normal distribution (see Figure 8.5).

FIGURE 8.5

Standardized normal distribution and *t* distribution for 5 degrees of freedom



The degrees of freedom, $n - 1$, are directly related to the sample size, n . The concept of *degrees of freedom* is discussed further on page 281. As the sample size and degrees of freedom increase, S becomes a better estimate of σ , and the *t* distribution gradually approaches the standardized normal distribution, until the two are virtually identical. With a sample size of about 120 or more, S estimates σ closely enough so that there is little difference between the *t* and *Z* distributions.

As stated earlier, the *t* distribution assumes that the random variable X is normally distributed. In practice, however, when the sample size is large enough and the population is not very skewed, in most cases you can use the *t* distribution to estimate the population mean when σ is unknown. When dealing with a small sample size and a skewed population distribution, the confidence interval estimate may not provide a valid estimate of the population mean. To assess the assumption of normality, you can evaluate the shape of the sample data by constructing a histogram, stem-and-leaf display, boxplot, or normal probability plot. However, the ability of any of these graphs to help you evaluate normality is limited when you have a small sample size.

You find the critical values of *t* for the appropriate degrees of freedom from the table of the *t* distribution (see Table E.3). The columns of the table present the most commonly used cumulative probabilities and corresponding upper-tail areas. The rows of the table represent the degrees of freedom. The critical *t* values are found in the cells of the table. For example, with 99 degrees of freedom, if you want 95% confidence, you find the appropriate value of *t*, as shown in Table 8.1. The 95% confidence level means that 2.5% of the values (an area of

TABLE 8.1

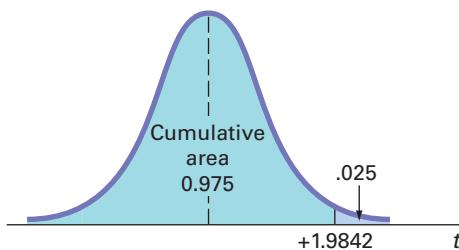
Determining the Critical Value from the *t* Table for an Area of 0.025 in Each Tail with 99 Degrees of Freedom

Degrees of Freedom	Cumulative Probabilities					
	.75	.90	.95	.975	.99	.995
	Upper-Tail Areas					
.25	.10	.05	.025	.01	.005	
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
⋮	⋮	⋮	⋮	⋮	⋮	⋮
96	0.6771	1.2904	1.6609	1.9850	2.3658	2.6280
97	0.6770	1.2903	1.6607	1.9847	2.3654	2.6275
98	0.6770	1.2902	1.6606	1.9845	2.3650	2.6269
99	0.6770	1.2902	1.6604	1.9842	2.3646	2.6264
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259

Source: Extracted from Table E.3.

0.025) are in each tail of the distribution. Looking in the column for a cumulative probability of 0.975 and an upper-tail area of 0.025 in the row corresponding to 99 degrees of freedom gives you a critical value for t of 1.9842 (see Figure 8.6). Because t is a symmetrical distribution with a mean of 0, if the upper-tail value is +1.9842, the value for the lower-tail area (lower 0.025) is -1.9842. A t value of -1.9842 means that the probability that t is less than -1.9842 is 0.025, or 2.5%.

FIGURE 8.6
 t distribution with
99 degrees of freedom



Note that for a 95% confidence interval, you will always have a cumulative probability of 0.975 and an upper-tail area of 0.025. Similarly, for a 99% confidence interval, you will have 0.995 and 0.005, and for a 90% confidence interval you will have 0.95 and 0.05.

The Concept of Degrees of Freedom

In Chapter 3, you learned that the numerator of the sample variance, S^2 [see Equation (3.6) on page 109], requires the computation of the sum of squares around the sample mean:

$$\sum_{i=1}^n (X_i - \bar{X})^2$$

In order to compute S^2 , you first need to know \bar{X} . Therefore, only $n - 1$ of the sample values are free to vary. This means that you have $n - 1$ degrees of freedom. For example, suppose a sample of five values has a mean of 20. How many values do you need to know before you can determine the remainder of the values? The fact that $n = 5$ and $\bar{X} = 20$ also tells you that

$$\sum_{i=1}^n X_i = 100$$

because

$$\frac{\sum_{i=1}^n X_i}{n} = \bar{X}$$

Thus, when you know four of the values, the fifth one is *not* free to vary because the sum must be 100. For example, if four of the values are 18, 24, 19, and 16, the fifth value must be 23, so that the sum is 100.

The Confidence Interval Statement

Equation (8.2) defines the $(1 - \alpha) \times 100\%$ confidence interval estimate for the mean with σ unknown.

CONFIDENCE INTERVAL FOR THE MEAN (σ UNKNOWN)

$$\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

or
$$\bar{X} - t_{\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2} \frac{S}{\sqrt{n}}$$

where

$t_{\alpha/2}$ is the critical value corresponding to an upper-tail probability of $\alpha/2$ (i.e., a cumulative area of $1 - \alpha/2$) from the t distribution with $n - 1$ degrees of freedom.

To illustrate the application of the confidence interval estimate for the mean when the standard deviation is unknown, recall the Ricknel Home Centers scenario presented on page 272. Using the DCOVA steps first discussed on page 2, you define the variable of interest as the dollar amount listed on the sales invoices for the month. Your business objective is to estimate the mean dollar amount. Then you collect the data by selecting a sample of 100 sales invoices from the population of sales invoices during the month. Once you have collected the data, you organize the data in a worksheet. You can construct various graphs (not shown here) to better visualize the distribution of the dollar amounts. To analyze the data, you compute the sample mean of the 100 sales invoices to be equal to \$110.27 and the sample standard deviation to be equal to \$28.95. For 95% confidence, the critical value from the t distribution (as shown in Table 8.1 on page 280) is 1.9842. Using Equation (8.2),

$$\begin{aligned} \bar{X} &\pm t_{\alpha/2} \frac{S}{\sqrt{n}} \\ &= 110.27 \pm (1.9842) \frac{28.95}{\sqrt{100}} \\ &= 110.27 \pm 5.74 \\ 104.53 &\leq \mu \leq 116.01 \end{aligned}$$

Figure 8.7 presents this confidence interval estimate of the mean dollar amount as computed by Excel and Minitab.

FIGURE 8.7

Excel and Minitab results for the confidence interval estimate for the mean sales invoice amount worksheet for the Ricknel Home Centers example

A	B	One-Sample T
1 Confidence Interval Estimate for the Mean		
2		
3 Data		
4 Sample Standard Deviation	28.95	N 100
5 Sample Mean	110.27	Mean 110.27
6 Sample Size	100	StDev 28.95
7 Confidence Level	95%	SE Mean 2.90
8		95% CI (104.53, 116.01)
9 Intermediate Calculations		
10 Standard Error of the Mean	2.895	=B4/SQRT(B6)
11 Degrees of Freedom	99	=B6 - 1
12 t Value	1.9842	=T.INV.2T(1 - B7, B11)
13 Interval Half Width	5.7443	=B12 * B10
14		
15 Confidence Interval		
16 Interval Lower Limit	104.53	=B5 - B13
17 Interval Upper Limit	116.01	=B5 + B13

Thus, with 95% confidence, you conclude that the mean amount of all the sales invoices is between \$104.53 and \$116.01. The 95% confidence level indicates that if you selected all possible samples of 100 (something that is never done in practice), 95% of the intervals developed would include the population mean somewhere within the interval. The validity of this confidence

interval estimate depends on the assumption of normality for the distribution of the amount of the sales invoices. With a sample of 100, the normality assumption is not overly restrictive, and the use of the t distribution is likely appropriate. Example 8.3 further illustrates how you construct the confidence interval for a mean when the population standard deviation is unknown.

EXAMPLE 8.3

Estimating the Mean Processing Time of Life Insurance Applications

TABLE 8.2

Processing Time for Life Insurance Applications

An insurance company has the business objective of reducing the amount of time it takes to approve applications for life insurance. The approval process consists of underwriting, which includes a review of the application, a medical information bureau check, possible requests for additional medical information and medical exams, and a policy compilation stage in which the policy pages are generated and sent for delivery. Using the DCOVA steps first discussed on page 2, you define the variable of interest as the total processing time in days. You collect the data by selecting a random sample of 27 approved policies during a period of one month. You organize the data collected in a worksheet. Table 8.2 lists the total processing time, in days, which are stored in **Insurance**. To analyze the data, you need to construct a 95% confidence interval estimate for the population mean processing time.

73	19	16	64	28	28	31	90	60	56	31	56	22	18
45	48	17	17	17	91	92	63	50	51	69	16	17	

SOLUTION To visualize the data, you construct a boxplot of the processing time, as displayed in Figure 8.8, and a normal probability plot, as shown in Figure 8.9. To analyze the data, you construct the confidence interval estimate shown in Figure 8.10.

FIGURE 8.8

Excel and Minitab boxplots for the processing time for life insurance applications

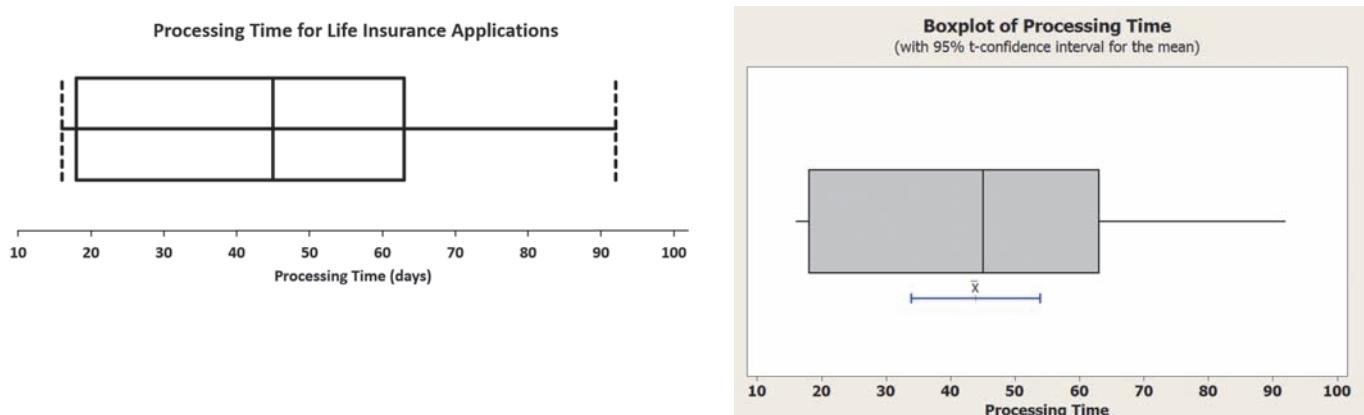
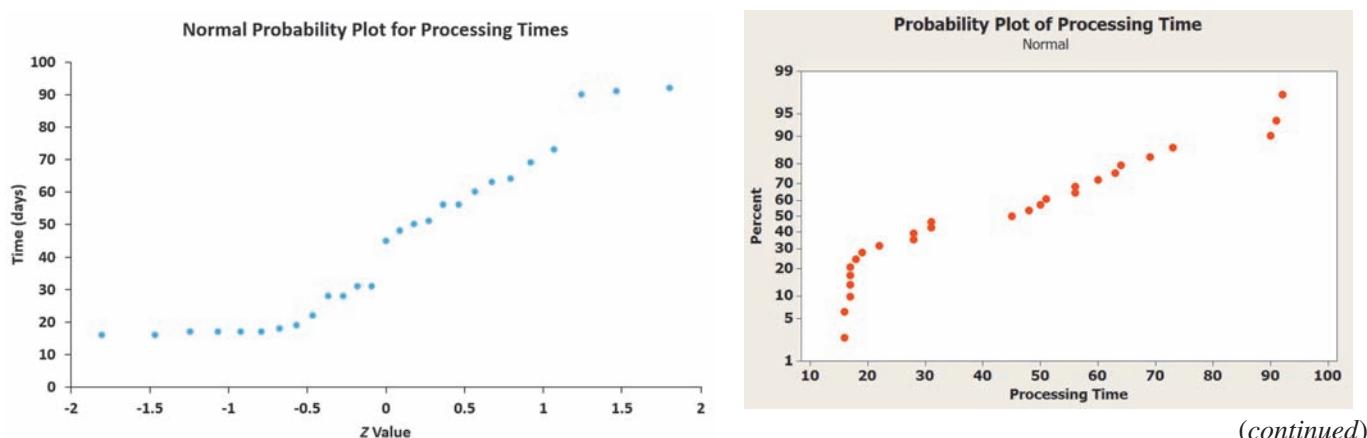


FIGURE 8.9

Excel and Minitab normal probability plots for the processing time for life insurance applications



(continued)

FIGURE 8.10

Excel and Minitab confidence interval estimates for the mean processing time worksheet for life insurance applications

A	B
Processing Time for Life Insurance Applications	
2	
Data	
4 Sample Standard Deviation	25.28
5 Sample Mean	43.89
6 Sample Size	27
7 Confidence Level	95%
Intermediate Calculations	
10 Standard Error of the Mean	4.8651
11 Degrees of Freedom	26
12 t Value	2.0555
13 Interval Half Width	10.0004
Confidence Interval	
16 Interval Lower Limit	33.89
17 Interval Upper Limit	53.89

One-Sample T: Time					
Variable	N	Mean	StDev	SE Mean	95% CI
Time	27	43.89	25.28	4.87	(33.89, 53.89)

Figure 8.10 shows that the sample mean is $\bar{X} = 43.89$ days and the sample standard deviation is $S = 25.28$ days. Using Equation (8.2) on page 282 to construct the confidence interval, you need to determine the critical value from the t table, using the row for 26 degrees of freedom. For 95% confidence, you use the column corresponding to an upper-tail area of 0.025 and a cumulative probability of 0.975. From Table E.3, you see that $t_{\alpha/2} = 2.0555$. Thus, using $\bar{X} = 43.89$, $S = 25.28$, $n = 27$, and $t_{\alpha/2} = 2.0555$,

$$\begin{aligned}\bar{X} &\pm t_{\alpha/2} \frac{S}{\sqrt{n}} \\ &= 43.89 \pm (2.0555) \frac{25.28}{\sqrt{27}} \\ &= 43.89 \pm 10.00 \\ 33.89 &\leq \mu \leq 53.89\end{aligned}$$

You conclude with 95% confidence that the mean processing time for the population of life insurance applications is between 33.89 and 53.89 days. The validity of this confidence interval estimate depends on the assumption that the processing time is normally distributed. From the boxplot displayed in Figure 8.8 and the normal probability plot shown in Figure 8.9, the processing time appears right-skewed. Thus, although the sample size is close to 30, you would have some concern about the validity of this confidence interval in estimating the population mean processing time. The concern is that a 95% confidence interval based on a small sample from a skewed distribution will contain the population mean less than 95% of the time in repeated sampling. In the case of small sample sizes and skewed distributions, you might consider the sample median as an estimate of central tendency and construct a confidence interval for the population median (see reference 2).

The interpretation of the confidence interval when σ is unknown is the same as when σ is known. To illustrate the fact that the confidence interval for the mean varies more when σ is unknown, return to the example concerning the order-filling times discussed in Section 8.1 on pages 275 and 276. Suppose that, in this case, you do *not* know the population standard deviation and instead use the sample standard deviation to construct the confidence interval estimate of the mean. Figure 8.11 shows the results for each of 20 samples of $n = 10$ orders.

In Figure 8.11 on page 285, observe that the standard deviation of the samples varies from 6.25 (sample 17) to 14.83 (sample 3). Thus, the width of the confidence interval developed varies from 8.94 in sample 17 to 21.22 in sample 3. Because you know that the population mean order time $\mu = 69.637$ minutes, you can see that the interval for sample 8 (69.68 – 85.48)

FIGURE 8.11

Confidence interval estimates of the mean for 20 samples of $n = 10$ randomly selected from the population of $N = 200$ orders with σ unknown

Sample	<i>N</i>	Mean	Std Dev	SE Mean	95% Conf. Int.
S01	10	71.64	7.58	2.40	(66.22, 77.06)
S02	10	67.22	10.95	3.46	(59.39, 75.05)
S03	10	67.97	14.83	4.69	(57.36, 78.58)
S04	10	73.90	10.59	3.35	(66.33, 81.47)
S05	10	67.11	11.12	3.52	(59.15, 75.07)
S06	10	68.12	10.83	3.43	(60.37, 75.87)
S07	10	65.80	10.85	3.43	(58.03, 73.57)
S08	10	77.58	11.04	3.49	(69.68, 85.48)
S09	10	66.69	11.45	3.62	(58.50, 74.88)
S10	10	62.55	8.58	2.71	(56.41, 68.69)
S11	10	71.12	12.82	4.05	(61.95, 80.29)
S12	10	70.55	10.52	3.33	(63.02, 78.08)
S13	10	65.51	8.16	2.58	(59.67, 71.35)
S14	10	64.90	7.55	2.39	(59.50, 70.30)
S15	10	66.22	11.21	3.54	(58.20, 74.24)
S16	10	70.43	10.21	3.23	(63.12, 77.74)
S17	10	72.04	6.25	1.96	(67.57, 76.51)
S18	10	73.91	11.29	3.57	(65.83, 81.99)
S19	10	71.49	9.76	3.09	(64.51, 78.47)
S20	10	70.15	10.84	3.43	(62.39, 77.91)

and the interval for sample 10 (56.41 – 68.69) do not correctly estimate the population mean. All the other intervals correctly estimate the population mean. Once again, remember that in practice you select only one sample, and you are unable to know for sure whether your one sample provides a confidence interval that includes the population mean.

Problems for Section 8.2

LEARNING THE BASICS

8.11 If $\bar{X} = 75$, $S = 24$, and $n = 36$, and assuming that the population is normally distributed, construct a 95% confidence interval estimate for the population mean, μ .

8.12 Determine the critical value of t in each of the following circumstances:

- a. $1 - \alpha = 0.95$, $n = 10$
- b. $1 - \alpha = 0.99$, $n = 10$
- c. $1 - \alpha = 0.95$, $n = 32$
- d. $1 - \alpha = 0.95$, $n = 65$
- e. $1 - \alpha = 0.90$, $n = 16$

8.13 Assuming that the population is normally distributed, construct a 95% confidence interval estimate for the population mean for each of the following samples:

Sample A: 1 1 1 1 8 8 8 8

Sample B: 1 2 3 4 5 6 7 8

Explain why these two samples produce different confidence intervals even though they have the same mean and range.

8.14 Assuming that the population is normally distributed, construct a 95% confidence interval for the population mean, based on the following sample of size $n = 7$:

1 2 3 4 5 6 20

Change the value of 20 to 7 and recalculate the confidence interval. Using these results, describe the effect of an outlier (i.e., an extreme value) on the confidence interval.

APPLYING THE CONCEPTS

8.15 A marketing researcher wants to estimate the mean savings (\$) realized by shoppers who showroom. Showrooming is the practice of inspecting products in retail stores and then purchasing the products online at a lower price. A random sample of 100 shoppers who recently purchased a consumer electronics item online after making a visit to a retail store yielded a mean savings of \$58 and a standard deviation of \$55.

- a. Construct a 95% confidence interval estimate for the mean savings for all showroomers who purchased a consumer electronics item.
- b. Suppose the owners of a consumer electronics retailer wants to estimate the total value of lost sales attributed to the next 1,000 showroomers that enter their retail store. How are the results in (a) useful in assisting the consumer electronics retailer in their estimation?



8.16 A survey of nonprofit organizations showed that online fundraising has increased in the past year. Based on a random sample of 55 nonprofits, the mean one-time gift donation in the past year was \$75, with a standard deviation of \$9.

- a. Construct a 95% confidence interval estimate for the population mean one-time gift donation.
- b. Interpret the interval constructed in (a).

8.17 The U.S. Department of Transportation requires tire manufacturers to provide tire performance information on the sidewall of a tire to better inform prospective customers as they make purchasing decisions. One very important measure of tire performance is the tread wear index, which indicates the tire's resistance to tread wear compared with a tire graded with a base of 100. A

tire with a grade of 200 should last twice as long, on average, as a tire graded with a base of 100. A consumer organization wants to estimate the actual tread wear index of a brand name of tires that claims “graded 200” on the sidewall of the tire. A random sample of $n = 18$ indicates a sample mean tread wear index of 195.3 and a sample standard deviation of 21.4.

- Assuming that the population of tread wear indexes is normally distributed, construct a 95% confidence interval estimate for the population mean tread wear index for tires produced by this manufacturer under this brand name.
- Do you think that the consumer organization should accuse the manufacturer of producing tires that do not meet the performance information provided on the sidewall of the tire? Explain.
- Explain why an observed tread wear index of 210 for a particular tire is not unusual, even though it is outside the confidence interval developed in (a).

8.18 The file **FastFood** contains the amount that a sample of 15 customers spent for lunch (\$) at a fast-food restaurant:

7.42 6.29 5.83 6.50 8.34 9.51 7.10 6.80 5.90
4.89 6.50 5.52 7.90 8.30 9.60

- Construct a 95% confidence interval estimate for the population mean amount spent for lunch (\$) at a fast-food restaurant, assuming a normal distribution.
- Interpret the interval constructed in (a).

8.19 The file **Sedans** contains the overall miles per gallon (MPG) of 2013 midsized sedans:

38 26 30 26 25 27 22 27 39 24 24 26 25
23 25 26 31 26 37 22 29 25 33 21 21

Source: Data extracted from “Ratings,” *Consumer Reports*, April 2013, pp. 30–31.

- Construct a 95% confidence interval estimate for the population mean MPG of 2013 family sedans, assuming a normal distribution.
- Interpret the interval constructed in (a).
- Compare the results in (a) to those in Problem 8.20(a).

8.20 The file **SUV** contains the overall MPG of 2013 small SUVs:

22 23 21 22 25 26 22 22 21
19 22 22 26 23 24 21 22

Source: Data extracted from “Ratings,” *Consumer Reports*, April 2013, pp. 34–35.

- Construct a 95% confidence interval estimate for the population mean MPG of 2013 small SUVs, assuming a normal distribution.
- Interpret the interval constructed in (a).
- Compare the results in (a) to those in Problem 8.19(a).

8.21 Is there a difference in the yields of different types of investments? The file **CDRate** contains the yields for a one-year certificate of deposit (CD) and a five-year CD for 23 banks in the United States as of March 20, 2013. (Data extracted from www.Bankrate.com, March 20, 2013.)

- Construct a 95% confidence interval estimate for the mean yield of one-year CDs.

- Construct a 95% confidence interval estimate for the mean yield of five-year CDs.
- Compare the results of (a) and (b).

8.22 One of the major measures of the quality of service provided by any organization is the speed with which the organization responds to customer complaints. A large family-held department store selling furniture and flooring, including carpet, had undergone a major expansion in the past several years. In particular, the flooring department had expanded from 2 installation crews to an installation supervisor, a measurer, and 15 installation crews. The store had the business objective of improving its response to complaints. The variable of interest was defined as the number of days between when the complaint was made and when it was resolved. Data were collected from 50 complaints that were made in the past year. The data, stored in **Furniture**, are as follows:

54	5	35	137	31	27	152	2	123	81	74	27
11	19	126	110	110	29	61	35	94	31	26	5
12	4	165	32	29	28	29	26	25	1	14	13
13	10	5	27	4	52	30	22	36	26	20	23
						33	68				

- Construct a 95% confidence interval estimate for the population mean number of days between the receipt of a complaint and the resolution of the complaint.
- What assumption must you make about the population distribution in order to construct the confidence interval estimate in (a)?
- Do you think that the assumption needed in order to construct the confidence interval estimate in (a) is valid? Explain.
- What effect might your conclusion in (c) have on the validity of the results in (a)?

8.23 A manufacturing company produces electric insulators. You define the variable of interest as the strength of the insulators. If the insulators break when in use, a short circuit is likely. To test the strength of the insulators, you carry out destructive testing to determine how much force is required to break the insulators. You measure force by observing how many pounds are applied to the insulator before it breaks. You collect the force data for 30 insulators selected for the experiment and organize and store these data in **Force**:

1,870 1,728 1,656 1,610 1,634 1,784 1,522 1,696
1,592 1,662 1,866 1,764 1,734 1,662 1,734 1,774
1,550 1,756 1,762 1,866 1,820 1,744 1,788 1,688
1,810 1,752 1,680 1,810 1,652 1,736

- Construct a 95% confidence interval estimate for the population mean force.
- What assumption must you make about the population distribution in order to construct the confidence interval estimate in (a)?
- Do you think that the assumption needed in order to construct the confidence interval estimate in (a) is valid? Explain.

8.24 The file **MarketPenetration** contains Facebook penetration values (the percentage of a country’s population that are Facebook users) for 15 countries:

52.56 33.09 5.37 19.41 32.52 41.69 51.61 30.12
39.07 30.62 38.16 49.35 27.13 53.45 40.01

Source: Data extracted from www.socialbakers.com/facebook-statistics/.

- Construct a 95% confidence interval estimate for the population mean Facebook penetration.
- What assumption do you need to make about the population to construct the interval in (a)?
- Given the data presented, do you think the assumption needed in (a) is valid? Explain.

8.25 One operation of a mill is to cut pieces of steel into parts that are used in the frame for front seats in an automobile. The steel is cut with a diamond saw, and the resulting parts must be cut to be within ± 0.005 inch of the length specified by the automobile company. The measurement reported from a sample of 100 steel parts (stored in **Steel**) is the difference, in inches, between the actual length of the steel part, as measured by a laser

measurement device, and the specified length of the steel part. For example, the first observation, -0.002 , represents a steel part that is 0.002 inch shorter than the specified length.

- Construct a 95% confidence interval estimate for the population mean difference between the actual length of the steel part and the specified length of the steel part.
- What assumption must you make about the population distribution in order to construct the confidence interval estimate in (a)?
- Do you think that the assumption needed in order to construct the confidence interval estimate in (a) is valid? Explain.
- Compare the conclusions reached in (a) with those of Problem 2.43 on page 64.

8.3 Confidence Interval Estimate for the Proportion

Student Tip

As noted in Chapter 7, do not confuse this use of the Greek letter pi, π , to represent the population proportion with the mathematical constant that is the ratio of the circumference to a diameter of a circle—approximately 3.14159—which is also known by the same Greek letter.

The concept of a confidence interval also applies to categorical data. With categorical data, you want to estimate the proportion of items in a population having a certain characteristic of interest. The unknown population proportion is represented by the Greek letter π . The point estimate for π is the sample proportion, $p = X/n$, where n is the sample size and X is the number of items in the sample having the characteristic of interest. Equation (8.3) defines the confidence interval estimate for the population proportion.

CONFIDENCE INTERVAL ESTIMATE FOR THE PROPORTION

$$p \pm Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

or

$$p - Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \quad (8.3)$$

where

$$p = \text{sample proportion} = \frac{X}{n} = \frac{\text{Number of items having the characteristic}}{\text{sample size}}$$

π = population proportion

$Z_{\alpha/2}$ = critical value from the standardized normal distribution

n = sample size

Note: To use this equation for the confidence interval, the sample size n must be large enough to ensure that both X and $n - X$ are greater than 5.

You can use the confidence interval estimate for the proportion defined in Equation (8.3) to estimate the proportion of sales invoices that contain errors (see the Ricknel Home Centers scenario on page 272). Using the DCOVA steps, you define the variable of interest as whether the invoice contains errors (yes or no). Then, you collect the data from a sample of 100 sales invoices. The results, which you organize and store in a worksheet, show that 10 invoices contain errors. To analyze the data, you compute, for these data, $p = X/n = 10/100 = 0.10$. Since both $X = 10$ and $n - X = 100 - 10 = 90$ are > 5 , using Equation (8.3) and $Z_{\alpha/2} = 1.96$, for 95% confidence,

$$\begin{aligned}
 p &\pm Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \\
 &= 0.10 \pm (1.96) \sqrt{\frac{(0.10)(0.90)}{100}} \\
 &= 0.10 \pm (1.96)(0.03) \\
 &= 0.10 \pm 0.0588 \\
 0.0412 &\leq \pi \leq 0.1588
 \end{aligned}$$

Therefore, you have 95% confidence that the population proportion of all sales invoices containing errors is between 0.0412 and 0.1588. This means that between 4.12% and 15.88% of all the sales invoices contain errors. Figure 8.12 shows a confidence interval estimate for this example.

FIGURE 8.12

Excel and Minitab confidence interval estimates for the proportion of sales invoices that contain errors worksheet

A	B
Proportion of In-Error Sales Invoices	
1	
Data	
4	Sample Size
5	Number of Successes
6	Confidence Level
Intermediate Calculations	
9	Sample Proportion
10	Z Value
11	Standard Error of the Proportion
12	Interval Half Width
Confidence Interval	
15	Interval Lower Limit
16	Interval Upper Limit

Test and CI for One Proportion				
Sample	X	N	Sample p	95% CI
1	10	100	0.100000	(0.041201, 0.158799)

Using the normal approximation.

Example 8.4 illustrates another application of a confidence interval estimate for the proportion.

EXAMPLE 8.4

Estimating the Proportion of Nonconforming Newspapers Printed

The operations manager at a large newspaper wants to estimate the proportion of newspapers printed that have a nonconforming attribute. Using the DCOVA steps, you define the variable of interest as whether the newspaper has excessive ruboff, improper page setup, missing pages, or duplicate pages. You collect the data by selecting a random sample of $n = 200$ newspapers from all the newspapers printed during a single day. You organize the results in a worksheet, which shows that 35 newspapers contain some type of nonconformance. To analyze the data, you need to construct and interpret a 90% confidence interval estimate for the proportion of newspapers printed during the day that have a nonconforming attribute.

SOLUTION Using Equation (8.3),

$$\begin{aligned}
 p &= \frac{X}{n} = \frac{35}{200} = 0.175, \text{ and with a 90\% level of confidence } Z_{\alpha/2} = 1.645 \\
 p &\pm Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \\
 &= 0.175 \pm (1.645) \sqrt{\frac{(0.175)(0.825)}{200}} \\
 &= 0.175 \pm (1.645)(0.0269) \\
 &= 0.175 \pm 0.0442 \\
 0.1308 &\leq \pi \leq 0.2192
 \end{aligned}$$

You conclude with 90% confidence that the population proportion of all newspapers printed that day with nonconformities is between 0.1308 and 0.2192. This means you estimate that between 13.08% and 21.92% of the newspapers printed on that day have some type of nonconformance.

Equation (8.3) contains a Z statistic because you can use the normal distribution to approximate the binomial distribution when the sample size is sufficiently large. In Example 8.4, the confidence interval using Z provides an excellent approximation for the population proportion because both X and $n - X$ are greater than 5. However, if you do not have a sufficiently large sample size, you should use the binomial distribution rather than Equation (8.3) (see references 1, 3, and 9). The exact confidence intervals for various sample sizes and proportions of items of interest have been tabulated by Fisher and Yates (reference 3) and can also be computed using Minitab.

Problems for Section 8.3

LEARNING THE BASICS

8.26 If $n = 200$ and $X = 50$, construct a 95% confidence interval estimate for the population proportion.

8.27 If $n = 400$ and $X = 25$, construct a 99% confidence interval estimate for the population proportion.

APPLYING THE CONCEPTS

 **8.28** A cellphone provider has the business objective of wanting to estimate the proportion of subscribers who would upgrade to a new cellphone with improved features if it were made available at a substantially reduced cost. Data are collected from a random sample of 500 subscribers. The results indicate that 135 of the subscribers would upgrade to a new cellphone at a reduced cost.

- Construct a 99% confidence interval estimate for the population proportion of subscribers that would upgrade to a new cellphone at a reduced cost.
- How would the manager in charge of promotional programs use the results in (a)?

8.29 In a survey of 3,773 travelers, 1,509 said that location was very important for choosing a hotel and 1,207 said that reputation was very important in choosing an airline. (Data extracted from “Travelers Get Stingy with Their Loyalty,” *USA Today*, January 18, 2013, p. 8B.)

- Construct a 95% confidence interval estimate for the population proportion of travelers who said that location was very important for choosing a hotel.
- Construct a 95% confidence interval estimate for the population proportion of travelers who said that reputation was very important in choosing an airline.
- Write a short summary of the information derived from (a) and (b).

8.30 Are you more likely to purchase a brand mentioned by an athlete on a social media site? According to a Catalyst Digital Fan Engagement survey, 53% of social media sports fans would make such a purchase. (Data extracted from “Survey: Social Media Continues to Fuel Fans,” *Sports Business Journal*, July 16, 2012, p. 24.)

- Suppose that the survey had a sample size of $n = 500$. Construct a 95% confidence interval estimate for the population proportion of social media sports fans that would more likely purchase a brand mentioned by an athlete on a social media site.
- Based on (a), can you claim that more than half of all social media sports fans would more likely purchase a brand mentioned by an athlete on a social media site?

- Repeat parts (a) and (b), assuming that the survey had a sample size of $n = 5,000$.

- Discuss the effect of sample size on confidence interval estimation.

8.31 In a survey of 280 qualified readers of *Logistics Management*, 62 responded that the “cloud” and Software as a Service (SaaS) is not an option for their firms, citing issues such as security and privacy concerns, system reliability and system performance, data integrity, and lack of control as the biggest concerns. (Data extracted from “2012 Supply Chain Software Users Survey: Spending Stabilizers,” *Logistics Management*, May 2012, p. 38.) Construct a 95% confidence interval estimate for the population proportion of logistics firms for which the cloud and SaaS is not an option.

8.32 In a survey of 1,954 cellphone owners, adults aged 18 and over, 743 reported that they use their phone to keep themselves occupied during commercials or breaks in something they were watching on television, while 430 used their phone to check whether something they heard on television is true. (Data extracted from “The Rise of the Connected Viewer,” Pew Research Center’s Internet & American Life Project, July 17, 2012, pewinternet.org/~/media//Files/Reports/2012/PIP_Connected_Viewers.pdf.)

- Construct a 95% confidence interval estimate for the population proportion of adult cellphone owners who report that they use their phone to keep themselves occupied during commercials or breaks in something they were watching on television.
- Construct a 95% confidence interval estimate for the population proportion of adult cellphone owners who report that they use their phone to check whether something they heard on television was true.
- Compare the results of (a) and (b).

8.33 What are the factors that influence technology (tech) CEOs’ anticipated need to change strategy? In a survey by PricewaterhouseCoopers (PwC), 94 of 115 tech CEOs around the globe responded that customer demand is one of the reasons they are making strategic changes at their organization, and 40 responded that availability of talent is one of the reasons. (Data extracted from “Delivering Results: Key Findings in the Technology Sector,” 15th Annual PwC Global CEO Survey, 2012.)

- Construct a 95% confidence interval estimate for the population proportion of tech CEOs who indicate customer demand as one of the reasons for making strategic change.
- Construct a 95% confidence interval estimate for the population proportion of tech CEOs who indicate availability of talent as one of the reasons for making strategic change.
- Interpret the intervals in (a) and (b).

8.4 Determining Sample Size

In each confidence interval developed so far in this chapter, the sample size was reported along with the results, with little discussion of the width of the resulting confidence interval. In the business world, sample sizes are determined prior to data collection to ensure that the confidence interval is narrow enough to be useful in making decisions. Determining the proper sample size is a complicated procedure, subject to the constraints of budget, time, and the amount of acceptable sampling error. In the Ricknel Home Centers scenario, if you want to estimate the mean dollar amount of the sales invoices, you must determine in advance how large a sampling error to allow in estimating the population mean. You must also determine, in advance, the level of confidence (i.e., 90%, 95%, or 99%) to use in estimating the population parameter.

Sample Size Determination for the Mean

To develop an equation for determining the appropriate sample size needed when constructing a confidence interval estimate for the mean, recall Equation (8.1) on page 276:

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

²In this context, some statisticians refer to e as the **margin of error**.

The amount added to or subtracted from \bar{X} is equal to half the width of the interval. This quantity represents the amount of imprecision in the estimate that results from sampling error.² The sampling error, e , is defined as

$$e = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Solving for n gives the sample size needed to construct the appropriate confidence interval estimate for the mean. “Appropriate” means that the resulting interval will have an acceptable amount of sampling error.

SAMPLE SIZE DETERMINATION FOR THE MEAN

The sample size, n , is equal to the product of the $Z_{\alpha/2}$ value squared and the standard deviation, σ , squared, divided by the square of the sampling error, e .

$$n = \frac{Z_{\alpha/2}^2 \sigma^2}{e^2} \tag{8.4}$$

To compute the sample size, you must know three quantities:

- The desired confidence level, which determines the value of $Z_{\alpha/2}$, the critical value from the standardized normal distribution³
- The acceptable sampling error, e
- The standard deviation, σ

In some business-to-business relationships that require estimation of important parameters, legal contracts specify acceptable levels of sampling error and the confidence level required. For companies in the food and drug sectors, government regulations often specify sampling errors and confidence levels. In general, however, it is usually not easy to specify the three quantities needed to determine the sample size. How can you determine the level of confidence and sampling error? Typically, these questions are answered only by a subject matter expert (i.e., an individual very familiar with the variables under study). Although 95% is the most common confidence level used, if more confidence is desired, then 99% might be more appropriate; if

³You use Z instead of t because, to determine the critical value of t , you need to know the sample size, but you do not know it yet. For most studies, the sample size needed is large enough that the standardized normal distribution is a good approximation of the t distribution.

less confidence is deemed acceptable, then 90% might be used. For the sampling error, you should think not of how much sampling error you would like to have (you really do not want any error) but of how much you can tolerate when reaching conclusions from the confidence interval.

In addition to specifying the confidence level and the sampling error, you need an estimate of the standard deviation. Unfortunately, you rarely know the population standard deviation, σ . In some instances, you can estimate the standard deviation from past data. In other situations, you can make an educated guess by taking into account the range and distribution of the variable. For example, if you assume a normal distribution, the range is approximately equal to 6σ (i.e., $\pm 3\sigma$ around the mean) so that you estimate σ as the range divided by 6. If you cannot estimate σ in this way, you can conduct a small-scale study and estimate the standard deviation from the resulting data.

To explore how to determine the sample size needed for estimating the population mean, consider again the audit at Ricknel Home Centers. In Section 8.2, you selected a sample of 100 sales invoices and constructed a 95% confidence interval estimate for the population mean sales invoice amount. How was this sample size determined? Should you have selected a different sample size?

Suppose that, after consulting with company officials, you determine that a sampling error of no more than $\pm \$5$ is desired, along with 95% confidence. Past data indicate that the standard deviation of the sales amount is approximately \$25. Thus, $e = \$5$, $\sigma = \$25$, and $Z_{\alpha/2} = 1.96$ (for 95% confidence). Using Equation (8.4),

$$n = \frac{Z_{\alpha/2}^2 \sigma^2}{e^2} = \frac{(1.96)^2(25)^2}{(5)^2} = 96.04$$

Because the general rule is to slightly oversatisfy the criteria by rounding the sample size up to the next whole integer, you should select a sample of size 97. Thus, the sample of size $n = 100$ used on page 282 is slightly more than what is necessary to satisfy the needs of the company, based on the estimated standard deviation, desired confidence level, and sampling error. Because the calculated sample standard deviation is slightly higher than expected, \$28.95 compared to \$25.00, the confidence interval is slightly wider than desired. Figure 8.13 shows a worksheet for determining the sample size.

FIGURE 8.13

Excel worksheet for determining the sample size for estimating the mean sales invoice amount for the Ricknel Home Centers example

A	B
1 For the Mean Sales Invoice Amount	
2	
3 Data	
4 Population Standard Deviation	25
5 Sampling Error	5
6 Confidence Level	95%
7	
8 Intermediate Calculations	
9 Z Value	-1.9600 =NORM.S.INV((1 - B6)/2)
10 Calculated Sample Size	96.0365 =((B9 * B4)/B5)^2
11	
12 Result	
13 Sample Size Needed	97 =ROUNDUP(B10, 0)

Example 8.5 illustrates another application of determining the sample size needed to develop a confidence interval estimate for the mean.

EXAMPLE 8.5

Determining the Sample Size for the Mean

Returning to Example 8.3 on page 283, suppose you want to estimate, with 95% confidence, the population mean processing time to within ± 4 days. On the basis of a study conducted the previous year, you believe that the standard deviation is 25 days. Determine the sample size needed.

(continued)

SOLUTION Using Equation (8.4) on page 290 and $e = 4$, $\sigma = 25$, and $Z_{\alpha/2} = 1.96$ for 95% confidence,

$$\begin{aligned} n &= \frac{Z_{\alpha/2}^2 \sigma^2}{e^2} = \frac{(1.96)^2(25)^2}{(4)^2} \\ &= 150.06 \end{aligned}$$

Therefore, you should select a sample of 151 applications because the general rule for determining sample size is to always round up to the next integer value in order to slightly oversatisfy the criteria desired. An actual sampling error slightly larger than 4 will result if the sample standard deviation calculated in this sample of 151 is greater than 25 and slightly smaller if the sample standard deviation is less than 25.

Sample Size Determination for the Proportion

So far in this section, you have learned how to determine the sample size needed for estimating the population mean. Now suppose that you want to determine the sample size necessary for estimating a population proportion.

To determine the sample size needed to estimate a population proportion, π , you use a method similar to the method for a population mean. Recall that in developing the sample size for a confidence interval for the mean, the sampling error is defined by

$$e = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

When estimating a proportion, you replace σ with $\sqrt{\pi(1 - \pi)}$. Thus, the sampling error is

$$e = Z_{\alpha/2} \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Solving for n , you have the sample size necessary to develop a confidence interval estimate for a proportion.

SAMPLE SIZE DETERMINATION FOR THE PROPORTION

The sample size n is equal to the product of $Z_{\alpha/2}$ squared, the population proportion, π , and 1 minus the population proportion, π , divided by the square of the sampling error, e .

$$n = \frac{Z_{\alpha/2}^2 \pi(1 - \pi)}{e^2} \quad (8.5)$$

To determine the sample size, you must know three quantities:

- The desired confidence level, which determines the value of $Z_{\alpha/2}$, the critical value from the standardized normal distribution
- The acceptable sampling error (or margin of error), e
- The population proportion, π

In practice, selecting these quantities requires some planning. Once you determine the desired level of confidence, you can find the appropriate $Z_{\alpha/2}$ value from the standardized normal distribution. The sampling error, e , indicates the amount of error that you are willing to tolerate in estimating the population proportion. The third quantity, π , is actually the

population parameter that you want to estimate! Thus, how do you state a value for what you are trying to determine?

Here you have two alternatives. In many situations, you may have past information or relevant experience that provides an educated estimate of π . If you do not have past information or relevant experience, you can try to provide a value for π that would never *underestimate* the sample size needed. Referring to Equation (8.5), you can see that the quantity $\pi(1 - \pi)$ appears in the numerator. Thus, you need to determine the value of π that will make the quantity $\pi(1 - \pi)$ as large as possible. When $\pi = 0.5$, the product $\pi(1 - \pi)$ achieves its maximum value. To show this result, consider the following values of π , along with the accompanying products of $\pi(1 - \pi)$:

$$\text{When } \pi = 0.9, \text{ then } \pi(1 - \pi) = (0.9)(0.1) = 0.09.$$

$$\text{When } \pi = 0.7, \text{ then } \pi(1 - \pi) = (0.7)(0.3) = 0.21.$$

$$\text{When } \pi = 0.5, \text{ then } \pi(1 - \pi) = (0.5)(0.5) = 0.25.$$

$$\text{When } \pi = 0.3, \text{ then } \pi(1 - \pi) = (0.3)(0.7) = 0.21.$$

$$\text{When } \pi = 0.1, \text{ then } \pi(1 - \pi) = (0.1)(0.9) = 0.09.$$

Therefore, when you have no prior knowledge or estimate for the population proportion, π , you should use $\pi = 0.5$ for determining the sample size. Using $\pi = 0.5$ produces the largest possible sample size and results in the narrowest and most precise confidence interval. This increased precision comes at the cost of spending more time and money for an increased sample size. Also, note that if you use $\pi = 0.5$ and the proportion is different from 0.5, you will overestimate the sample size needed, because you will get a confidence interval narrower than originally intended.

Returning to the Ricknel Home Centers scenario on page 272, suppose that the auditing procedures require you to have 95% confidence in estimating the population proportion of sales invoices with errors to within ± 0.07 . The results from past months indicate that the largest proportion has been no more than 0.15. Thus, using Equation (8.5) with $e = 0.07$, $\pi = 0.15$, and $Z_{\alpha/2} = 1.96$ for 95% confidence,

$$\begin{aligned} n &= \frac{Z_{\alpha/2}^2 \pi(1 - \pi)}{e^2} \\ &= \frac{(1.96)^2(0.15)(0.85)}{(0.07)^2} \\ &= 99.96 \end{aligned}$$

Because the general rule is to round the sample size up to the next whole integer to slightly oversatisfy the criteria, a sample size of 100 is needed. Thus, the sample size needed to satisfy the requirements of the company, based on the estimated proportion, desired confidence level, and sampling error, is equal to the sample size taken on page 287. The actual confidence interval is narrower than required because the sample proportion is 0.10, whereas 0.15 was used for π in Equation (8.5). Figure 8.14 shows a worksheet for determining the sample size.

FIGURE 8.14

Excel worksheet for determining sample size for estimating the proportion of in-error sales invoices for Ricknel Home Centers

A	B
1 For the Proportion of In-Error Sales Invoices	
2	
3 Data	
4 Estimate of True Proportion	0.15
5 Sampling Error	0.07
6 Confidence Level	95%
7	
8 Intermediate Calculations	
9 Z Value	-1.9600 =NORM.S.INV((1 - B6) / 2)
10 Calculated Sample Size	99.9563 =(B9^2 * B4 * (1 - B4)) / B5^2
11	
12 Result	
13 Sample Size Needed	100 =ROUNDUP(B10, 0)

Example 8.6 provides another application of determining the sample size for estimating the population proportion.

EXAMPLE 8.6

Determining the Sample Size for the Population Proportion

You want to have 90% confidence of estimating the proportion of office workers who respond to email within an hour to within ± 0.05 . Because you have not previously undertaken such a study, there is no information available from past data. Determine the sample size needed.

SOLUTION Because no information is available from past data, assume that $\pi = 0.50$. Using Equation (8.5) on page 292 and $e = 0.05$, $\pi = 0.50$, and $Z_{\alpha/2} = 1.645$ for 90% confidence,

$$\begin{aligned} n &= \frac{Z_{\alpha/2}^2 \pi(1 - \pi)}{e^2} \\ &= \frac{(1.645)^2(0.50)(0.50)}{(0.05)^2} \\ &= 270.6 \end{aligned}$$

Therefore, you need a sample of 271 office workers to estimate the population proportion to within ± 0.05 with 90% confidence.

Problems for Section 8.4

LEARNING THE BASICS

8.34 If you want to be 95% confident of estimating the population mean to within a sampling error of ± 5 and the standard deviation is assumed to be 15, what sample size is required?

8.35 If you want to be 99% confident of estimating the population mean to within a sampling error of ± 20 and the standard deviation is assumed to be 100, what sample size is required?

8.36 If you want to be 99% confident of estimating the population proportion to within a sampling error of ± 0.04 , what sample size is needed?

8.37 If you want to be 95% confident of estimating the population proportion to within a sampling error of ± 0.02 and there is historical evidence that the population proportion is approximately 0.40, what sample size is needed?

APPLYING THE CONCEPTS

8.38 A survey is planned to determine the mean annual family medical expenses of employees of a large company. The management of the company wishes to be 95% confident that the sample mean is correct to within $\pm \$50$ of the population mean annual family medical expenses. A previous study indicates that the standard deviation is approximately \$400.

- a. How large a sample is necessary?
- b. If management wants to be correct to within $\pm \$25$, how many employees need to be selected?

8.39 If the manager of a bottled water distributor wants to estimate, with 95% confidence, the mean amount of water in a 1-gallon bottle to within ± 0.004 gallon and also assumes that the standard deviation is 0.02 gallon, what sample size is needed?

8.40 If a light bulb manufacturing company wants to estimate, with 95% confidence, the mean life of compact fluorescent light bulbs to within ± 200 hours and also assumes that the population standard deviation is 1,000 hours, how many compact fluorescent light bulbs need to be selected?

8.41 If the inspection division of a county weights and measures department wants to estimate the mean amount of soft-drink fill in 2-liter bottles to within ± 0.01 liter with 95% confidence and also assumes that the standard deviation is 0.05 liter, what sample size is needed?

8.42 An advertising executive wants to estimate the mean weekly amount of time 18- to 24-year-olds spend watching traditional television in a large city. Based on studies in other cities, the standard deviation is assumed to be 10 minutes. The executive wants to estimate, with 99% confidence, the mean weekly amount of time to within ± 3 minutes.

- a. What sample size is needed?
- b. If 95% confidence is desired, how many 18- to 24-year-olds need to be selected?

8.43 An advertising agency that serves a major radio station wants to estimate the mean amount of time that the station's audience spends listening to the radio daily. From past studies, the standard deviation is estimated as 45 minutes.

- a. What sample size is needed if the agency wants to be 90% confident of being correct to within ± 5 minutes?
- b. If 99% confidence is desired, how many listeners need to be selected?

8.44 A growing niche in the restaurant business is gourmet-casual breakfast, lunch, and brunch. Chains in this group include EggSpec-tation and Panera Bread. Suppose that the mean per-person check

for breakfast at EggSpectation is approximately \$14.50, and the mean per-person check for Panera Bread is \$8.50.

- a. Assuming a standard deviation of \$2.00, what sample size is needed to estimate, with 95% confidence, the mean per-person check for EggSpectation to within $\pm \$0.25$?
- b. Assuming a standard deviation of \$2.50, what sample size is needed to estimate, with 95% confidence, the mean per-person check for EggSpectation to within $\pm \$0.25$?
- c. Assuming a standard deviation of \$3.00, what sample size is needed to estimate, with 95% confidence, the mean per-person check for EggSpectation to within $\pm \$0.25$?
- d. Discuss the effect of variation on the sample size needed.

8.45 What advertising medium is most influential in making a purchase decision? According to a TVB survey, 37.2% of American adults point to TV. (Data extracted from “TV Seen Most Influential Ad Medium for Purchase Decisions,” *MC Marketing Charts*, June 18, 2012.)

- a. To conduct a follow-up study that would provide 95% confidence that the point estimate is correct to within ± 0.04 of the population proportion, how large a sample size is required?
- b. To conduct a follow-up study that would provide 99% confidence that the point estimate is correct to within ± 0.04 of the population proportion, how many people need to be sampled?
- c. To conduct a follow-up study that would provide 95% confidence that the point estimate is correct to within ± 0.02 of the population proportion, how large a sample size is required?
- d. To conduct a follow-up study that would provide 99% confidence that the point estimate is correct to within ± 0.02 of the population proportion, how many people need to be sampled?
- e. Discuss the effects on sample size requirements of changing the desired confidence level and the acceptable sampling error.

8.46 A survey of 300 U.S. online shoppers was conducted. In response to the question of what would influence the shopper to spend more money online in 2012, 18% said free shipping, 13% said offering discounts while shopping, and 9% said product reviews. (Data extracted from “2012 Consumer Shopping Trends and Insights,” Steelhouse, Inc., 2012.) Construct a 95% confidence interval estimate of the population proportion of online shoppers who would be influenced to spend more money online in 2012 with

- a. free shipping.
- b. discounts offered while shopping.
- c. product reviews.

- d. You have been asked to update the results of this study. Determine the sample size necessary to estimate, with 95% confidence, the population proportions in (a) through (c) to within ± 0.02 .

8.47 In a study of 368 San Francisco Bay Area nonprofits, 224 reported that they are collaborating with other organizations to provide services, a necessity as nonprofit agencies are called upon to do more with less. (Data extracted from “2012 Nonprofit Pulse Survey,” United Way of the Bay Area, 2012, bit.ly/MkGINA.)

- a. Construct a 95% confidence interval for the proportion of San Francisco Bay Area nonprofits that collaborated with other organizations to provide services.
- b. Interpret the interval constructed in (a).
- c. If you wanted to conduct a follow-up study to estimate the population proportion of San Francisco Bay Area nonprofits that collaborated with other organizations to provide service to within ± 0.01 with 95% confidence, how may Bay Area nonprofits would you survey?

8.48 According to a new study released by Infosys, a global leader in consulting, outsourcing, and technology, more than three-quarters (77%) of U.S. consumers say that banking on their mobile device is convenient. (Data extracted from “Infosys Survey Finds Mobile Banking Customers Love Ease and Convenience, Yet Reliability and Security Concerns Remain,” *PR Newswire*, 2012, bit.ly/Ip9RUF.)

- a. If you conduct a follow-up study to estimate the population proportion of U.S. consumers who say that banking on their mobile device is convenient, would you use a π of 0.77 or 0.50 in the sample size formula?
- b. Using your answer in part (a), find the sample size necessary to estimate, with 95% confidence, the population proportion to within ± 0.03 .

8.49 Which store do you think is more expensive—physical or online? A recent survey (*USA Today*, December 10, 2012, p. 1B) found that 46% of people aged 20 to 40 thought that physical stores were more expensive.

- a. To conduct a follow-up study that would provide 99% confidence that the point estimate is correct to within ± 0.03 of the population proportion, how many people aged 20 to 40 need to be sampled?
- b. To conduct a follow-up study that would provide 99% confidence that the point estimate is correct to within ± 0.05 of the population proportion, how many people aged 20 to 40 need to be sampled?
- c. Compare the results of (a) and (b).

8.5 Confidence Interval Estimation and Ethical Issues

The selection of samples and the inferences that accompany them raise several ethical issues. The major ethical issue concerns whether confidence interval estimates accompany point estimates. Failure to include a confidence interval estimate might mislead the user of the results into thinking that the point estimate is all that is needed to predict the population characteristic with certainty. Confidence interval limits (typically set at 95%), the sample size used, and an interpretation of the meaning of the confidence interval in terms that a person untrained in statistics can understand should always accompany point estimates.

When media outlets publicize the results of a political poll, they often overlook including this type of information. Sometimes, the results of a poll include the sampling error, but the sampling error is often presented in fine print or as an afterthought to the story being reported.

A fully ethical presentation of poll results would give equal prominence to the confidence levels, sample size, sampling error, and confidence limits of the poll.

When you prepare your own point estimates, always state the interval estimate in a *prominent* place and include a brief explanation of the meaning of the confidence interval. In addition, make sure you highlight the sample size and sampling error.

8.6 Application of Confidence Interval Estimation in Auditing

Auditing is the collection and evaluation of evidence about information related to an economic entity in order to determine and report on how well the information corresponds to established criteria. Auditing uses probability sampling methods to develop confidence interval estimates. The **Section 8.6 online topic** reviews three common applications of confidence interval estimation in auditing.

8.7 Estimation and Sample Size Estimation for Finite Populations

To develop confidence interval estimates for population parameters or determine sample sizes when estimating population parameters, you use the finite population correction factor when samples are selected without replacement from a finite population. The **Section 8.7 online topic** explains how to use the finite population correction factor for these purposes.

8.8 Bootstrapping

The confidence interval estimation procedures discussed in this chapter make assumptions that are often not valid, especially for small samples. Bootstrapping, the selection of an initial sample and repeated sampling from that initial sample, provides an alternative approach that does not rely on those assumptions. The **Section 8.8 online topic** explains this alternative technique.

USING STATISTICS

Getting Estimates at Ricknel Home Centers, Revisited

In the Ricknel Home Centers scenario, you were an accountant for a distributor of home improvement supplies in the northeastern United States. You were responsible for the accuracy of the integrated inventory management and sales information system. You used confidence interval estimation techniques to draw conclusions about the population of all records from a relatively small sample collected during an audit.

At the end of the month, you collected a random sample of 100 sales invoices and made the following inferences:

- With 95% confidence, you concluded that the mean amount of all the sales invoices is between \$104.53 and \$116.01.



Mangostock/Shutterstock

- With 95% confidence, you concluded that between 4.12% and 15.88% of all the sales invoices contain errors.

These estimates provide an interval of values that you believe contain the true population parameters. If these intervals are too wide (i.e., the sampling error is too large) for the types of decisions Ricknel Home Centers needs to make, you will need to take a larger sample. You can use the sample size formulas in Section 8.4 to determine the number of sales invoices to sample to ensure that the size of the sampling error is acceptable.

SUMMARY

This chapter discusses confidence intervals for estimating the characteristics of a population, along with how you can determine the necessary sample size. You learned how to apply these methods to numerical and categorical data. Table 8.3 provides a list of topics covered in this chapter.

To determine what equation to use for a particular situation, you need to answer these questions:

TABLE 8.3

Summary of Topics in Chapter 8

TYPE OF ANALYSIS	TYPE OF DATA	
	Numerical	Categorical
Confidence interval for a population parameter	Confidence interval estimate for the mean (Sections 8.1 and 8.2)	Confidence interval estimate for the proportion (Section 8.3)
Determining sample size	Sample size determination for the mean (Section 8.4)	Sample size determination for the proportion (Section 8.4)

REFERENCES

1. Cochran, W. G. *Sampling Techniques*, 3rd ed. New York: Wiley, 1977.
2. Daniel, W. W. *Applied Nonparametric Statistics*, 2nd ed. Boston: PWS Kent, 1990.
3. Fisher, R. A., and F. Yates. *Statistical Tables for Biological, Agricultural and Medical Research*, 5th ed. Edinburgh: Oliver & Boyd, 1957.
4. Hahn, G., and W. Meeker. *Statistical Intervals: A Guide for Practitioners*. New York: John Wiley and Sons, Inc., 1991.
5. Kirk, R. E., ed. *Statistical Issues: A Reader for the Behavioral Sciences*. Belmont, CA: Wadsworth, 1972.
6. Larsen, R. L., and M. L. Marx. *An Introduction to Mathematical Statistics and Its Applications*, 4th ed. Upper Saddle River, NJ: Prentice Hall, 2006.
7. Microsoft Excel 2013. Redmond, WA: Microsoft Corp., 2012.
8. Minitab Release 16. State College, PA: Minitab, Inc., 2010.
9. Snedecor, G. W., and W. G. Cochran. *Statistical Methods*, 7th ed. Ames, IA: Iowa State University Press, 1980.

KEY EQUATIONS

Confidence Interval for the Mean (σ Known)

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

or

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (8.1)$$

Confidence Interval for the Mean (σ Unknown)

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

or

$$\bar{X} - t_{\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2} \frac{s}{\sqrt{n}} \quad (8.2)$$

Confidence Interval Estimate for the Proportion

$$p \pm Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

or

$$p - Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \quad (8.3)$$

Sample Size Determination for the Mean

$$n = \frac{Z_{\alpha/2}^2 \sigma^2}{e^2} \quad (8.4)$$

Sample Size Determination for the Proportion

$$n = \frac{Z_{\alpha/2}^2 \pi(1-\pi)}{e^2} \quad (8.5)$$

KEY TERMS

confidence interval estimate 273
 critical value 277
 degrees of freedom 279

level of confidence 276
 margin of error 290
 point estimate 273

sampling error 276
 Student's *t* distribution 279

CHECKING YOUR UNDERSTANDING

8.50 Why can you never really have 100% confidence of correctly estimating the population characteristic of interest?

8.51 When should you use the *t* distribution to develop the confidence interval estimate for the mean?

8.52 Why is it true that for a given sample size, n , an increase in confidence is achieved by widening (and making less precise) the confidence interval?

8.53 Why is the sample size needed to determine the proportion smaller when the population proportion is 0.20 than when the population proportion is 0.50?

CHAPTER REVIEW PROBLEMS

8.54 The Pew Internet Project survey of 2,253 American adults (data extracted from pewinternet.org/Commentary/2012/February/Pew-Internet-Mobile) found the following:

- 1,983 have a cellphone
 - 1,307 have a desktop computer
 - 1,374 have a laptop computer
 - 406 have an ebook reader
 - 406 have a tablet computer
- Construct 95% confidence interval estimates for the population proportion of the electronic devices adults own.
 - What conclusions can you reach concerning what electronic devices adults have?

8.55 What do Americans do to conserve energy? The Associated Press-NORC Center for Public Affairs Research conducted a survey of 897 adults who had personally done something to try to save energy in the last year (data extracted from "Energy Efficiency and Independence: How the Public Understands, Learns, and Acts," bit.ly/Maw5hd), and found the following percentages:

- Turn off lights: 39%
 - Turn down heat: 26%
 - Install more energy-saving appliances: 23%
 - Drive less/walk more/bicycle more: 18%
 - Unplug things: 16%
- Construct a 95% confidence interval estimate for the population proportion of what adults do to conserve energy.
 - What conclusions can you reach concerning what adults do to conserve energy?

8.56 A market researcher for a consumer electronics company wants to study the media viewing behavior of residents of a particular area. A random sample of 40 respondents is selected, and each respondent is instructed to keep a detailed record of time spent engaged viewing content across all screens (traditional TV, DVD/Blu-ray, game console, Internet on a computer, video on a

computer, video on a mobile phone) in a particular week. The results are as follows:

- Content viewing time per week: $\bar{X} = 41$ hours, $S = 3.5$ hours.
 - 30 respondents have high definition (HD) on at least one television set.
- Construct a 95% confidence interval estimate for the mean content viewing time per week in this area.
 - Construct a 95% confidence interval estimate for the population proportion of residents who have HD on at least one television set.
- Suppose that the market researcher wants to take another survey in a different location. Answer these questions:
- What sample size is required to be 95% confident of estimating the population mean content viewing time to within ± 2 hours assuming that the population standard deviation is equal to 5 hours?
 - How many respondents need to be selected to be 95% confident of being within ± 0.06 of the population proportion who have HD on at least one television set if no previous estimate is available?
 - Based on (c) and (d), how many respondents should the market researcher select if a single survey is being conducted?

8.57 An information technology (IT) consulting firm specializing in healthcare solutions wants to study communication deficiencies in the health care industry. A random sample of 70 health care clinicians reveals the following:

- Time wasted in a day due to outdated communication technologies: $\bar{X} = 45$ minutes, $S = 10$ minutes.
 - Thirty-six health care clinicians cite inefficiency of pagers as the reason for the wasted time.
- Construct a 99% confidence interval estimate for the population mean time wasted in a day due to outdated communication technologies.
 - Construct a 95% confidence interval estimate for the population proportion of health care clinicians who cite inefficiency of pagers as the reason for the wasted time.

8.58 The human resource (HR) director of a large corporation wishes to study absenteeism among its mid-level managers at its central office during the year. A random sample of 25 mid-level managers reveals the following:

- Absenteeism: $\bar{X} = 6.2$ days, $S = 7.3$ days.
 - 13 mid-level managers cite stress as a cause of absence.
- Construct a 95% confidence interval estimate for the mean number of absences for mid-level managers during the year.
 - Construct a 95% confidence interval estimate for the population proportion of mid-level managers who cite stress as a cause of absence.
- Suppose that the HR director wishes to administer a survey in one of its regional offices. Answer these questions:
- What sample size is needed to have 95% confidence in estimating the population mean absenteeism to within ± 1.5 days if the population standard deviation is estimated to be 8 days?
 - How many mid-level managers need to be selected to have 90% confidence in estimating the population proportion of mid-level managers who cite stress as a cause of absence to within ± 0.075 if no previous estimate is available?
 - Based on (c) and (d), what sample size is needed if a single survey is being conducted?

8.59 A national association devoted to human resource (HR) and workplace programs, practices, and training wants to study HR department practices and employee turnover of its member organizations. HR professionals and organization executives focus on turnover not only because it has significant cost implications but also because it affects overall business performance. A survey is designed to estimate the proportion of member organizations that have both talent and development programs in place to drive human-capital management as well as the member organizations' mean annual employee turnover rate (the ratio of the number of employees that left an organization in a given time period to the average number of employees in the organization during the given time period). A random sample of 100 member organizations reveals the following:

- Annual turnover rate: $\bar{X} = 8.1\%$, $S = 1.5\%$.
 - Thirty member organizations have both talent and development programs in place to drive human-capital management.
- Construct a 95% confidence interval estimate for the population mean annual turnover rate of member organizations.
 - Construct a 95% confidence interval estimate for the population proportion of member organizations that have both talent and development programs in place to drive human-capital management.
 - What sample size is needed to have 99% confidence of estimating the population mean annual employee turnover rate to within $\pm 1.5\%$?
 - How many member organizations need to be selected to have 90% confidence of estimating the population proportion of organizations that have both talent and development programs in place to drive human-capital management to within $\pm .045$?

8.60 The financial impact of IT systems downtime is a concern of plant operations management today. A survey of manufacturers examined the satisfaction level with the reliability and availability of their manufacturing IT applications. The variables of focus are whether the manufacturer experienced downtime in the past year that affected one or more manufacturing IT applications, the number of downtime incidents that occurred in the past year, and the

approximate cost of a typical downtime incident. The results from a sample of 200 manufacturers are as follows:

- Sixty-two experienced downtime this year that affected one or more manufacturing applications.
 - Number of downtime incidents: $\bar{X} = 3.5$, $S = 2.0$
 - Cost of downtime incidents: $\bar{X} = \$18,000$, $S = \$3,000$.
- Construct a 90% confidence interval estimate for the population proportion of manufacturers who experienced downtime in the past year that affected one or more manufacturing IT applications.
 - Construct a 95% confidence interval estimate for the population mean number of downtime incidents experienced by manufacturers in the past year.
 - Construct a 95% confidence interval estimate for the population mean cost of downtime incidents.

8.61 The branch manager of an outlet (Store 1) of a nationwide chain of pet supply stores wants to study characteristics of her customers. In particular, she decides to focus on two variables: the amount of money spent by customers and whether the customers own only one dog, only one cat, or more than one dog and/or cat. The results from a sample of 70 customers are as follows:

- Amount of money spent: $\bar{X} = \$21.34$, $S = \$9.22$.
 - Thirty-seven customers own only a dog.
 - Twenty-six customers own only a cat.
 - Seven customers own more than one dog and/or cat.
- Construct a 95% confidence interval estimate for the population mean amount spent in the pet supply store.
 - Construct a 90% confidence interval estimate for the population proportion of customers who own only a cat.

The branch manager of another outlet (Store 2) wishes to conduct a similar survey in his store. The manager does not have access to the information generated by the manager of Store 1. Answer the following questions:

- What sample size is needed to have 95% confidence of estimating the population mean amount spent in this store to within $\pm \$1.50$ if the standard deviation is estimated to be \$10?
- How many customers need to be selected to have 90% confidence of estimating the population proportion of customers who own only a cat to within ± 0.045 ?
- Based on your answers to (c) and (d), how large a sample should the manager take?

8.62 Scarlett and Heather, the owners of an upscale restaurant in Dayton, Ohio, want to study the dining characteristics of their customers. They decide to focus on two variables: the amount of money spent by customers and whether customers order dessert. The results from a sample of 60 customers are as follows:

- Amount spent: $\bar{X} = \$38.54$, $S = \$7.26$.
 - Eighteen customers purchased dessert.
- Construct a 95% confidence interval estimate for the population mean amount spent per customer in the restaurant.
 - Construct a 90% confidence interval estimate for the population proportion of customers who purchase dessert.
- Jeanine, the owner of a competing restaurant, wants to conduct a similar survey in her restaurant. Jeanine does not have access to the information that Scarlett and Heather have obtained from the survey they conducted. Answer the following questions:

- c. What sample size is needed to have 95% confidence of estimating the population mean amount spent in her restaurant to within $\pm \$1.50$, assuming that the standard deviation is estimated to be \$8?
- d. How many customers need to be selected to have 90% confidence of estimating the population proportion of customers who purchase dessert to within ± 0.04 ?
- e. Based on your answers to (c) and (d), how large a sample should Jeanine take?

8.63 The manufacturer of Ice Melt claims that its product will melt snow and ice at temperatures as low as 0° Fahrenheit. A representative for a large chain of hardware stores is interested in testing this claim. The chain purchases a large shipment of 5-pound bags for distribution. The representative wants to know, with 95% confidence and within ± 0.05 , what proportion of bags of Ice Melt perform the job as claimed by the manufacturer.

- a. How many bags does the representative need to test? What assumption should be made concerning the population proportion? (This is called *destructive testing*; i.e., the product being tested is destroyed by the test and is then unavailable to be sold.)
- b. Suppose that the representative tests 50 bags, and 42 of them do the job as claimed. Construct a 95% confidence interval estimate for the population proportion that will do the job as claimed.
- c. How can the representative use the results of (b) to determine whether to sell the Ice Melt product?

8.64 Claims fraud (illegitimate claims) and buildup (exaggerated loss amounts) continue to be major issues of concern among automobile insurance companies. Fraud is defined as specific material misrepresentation of the facts of a loss; buildup is defined as the inflation of an otherwise legitimate claim. A recent study examined auto injury claims closed with payment under private passenger coverages. Detailed data on injury, medical treatment, claimed losses, and total payments, as well as claim-handling techniques, were collected. In addition, auditors were asked to review the claim files to indicate whether specific elements of fraud or buildup appeared in the claim and, in the case of buildup, to specify the amount of excess payment. The file **InsuranceClaims** contains data for 90 randomly selected auto injury claims. The following variables are included: CLAIM—Claim ID; BUILDUP—1 if buildup indicated, 0 if not; and EXCESSPAYMENT—excess payment amount, in dollars.

- a. Construct a 95% confidence interval for the population proportion of all auto injury files that have exaggerated loss amounts.
- b. Construct a 95% confidence interval for the population mean dollar excess payment amount.

8.65 A quality characteristic of interest for a teabag-filling process is the weight of the tea in the individual bags. In this example, the label weight on the package indicates that the mean amount is 5.5 grams of tea in a bag. If the bags are underfilled, two problems arise. First, customers may not be able to brew the tea to be as strong as they wish. Second, the company may be in violation of the truth-in-labeling laws. On the other hand, if the mean amount of tea in a bag exceeds the label weight, the company is giving away product. Getting an exact amount of tea in a bag is problematic because of variation in the temperature and humidity inside the factory, differences in the density of the tea, and the extremely

fast filling operation of the machine (approximately 170 bags per minute). The following data (stored in **Teabags**) are the weights, in grams, of a sample of 50 tea bags produced in one hour by a single machine:

5.65	5.44	5.42	5.40	5.53	5.34	5.54	5.45	5.52	5.41
5.57	5.40	5.53	5.54	5.55	5.62	5.56	5.46	5.44	5.51
5.47	5.40	5.47	5.61	5.53	5.32	5.67	5.29	5.49	5.55
5.77	5.57	5.42	5.58	5.58	5.50	5.32	5.50	5.53	5.58
5.61	5.45	5.44	5.25	5.56	5.63	5.50	5.57	5.67	5.36

- a. Construct a 99% confidence interval estimate for the population mean weight of the tea bags.
- b. Is the company meeting the requirement set forth on the label that the mean amount of tea in a bag is 5.5 grams?
- c. Do you think the assumption needed to construct the confidence interval estimate in (a) is valid?

8.66 A manufacturing company produces steel housings for electrical equipment. The main component part of the housing is a steel trough that is made from a 14-gauge steel coil. It is produced using a 250-ton progressive punch press with a wipe-down operation that puts two 90-degree forms in the flat steel to make the trough. The distance from one side of the form to the other is critical because of weatherproofing in outdoor applications. The widths (in inches), shown below and stored in **Trough**, are from a sample of 49 troughs:

8.312	8.343	8.317	8.383	8.348	8.410	8.351	8.373	8.481
8.422	8.476	8.382	8.484	8.403	8.414	8.419	8.385	8.465
8.498	8.447	8.436	8.413	8.489	8.414	8.481	8.415	8.479
8.429	8.458	8.462	8.460	8.444	8.429	8.460	8.412	8.420
8.410	8.405	8.323	8.420	8.396	8.447	8.405	8.439	8.411
8.427	8.420	8.498	8.409					

- a. Construct a 95% confidence interval estimate for the mean width of the troughs.
- b. Interpret the interval developed in (a).
- c. Do you think the assumption needed to construct the confidence interval estimate in (a) is valid?

8.67 The manufacturer of Boston and Vermont asphalt shingles knows that product weight is a major factor in a customer's perception of quality. The last stage of the assembly line packages the shingles before they are placed on wooden pallets. Once a pallet is full (a pallet for most brands holds 16 squares of shingles), it is weighed, and the measurement is recorded. The file **Pallet** contains the weight (in pounds) from a sample of 368 pallets of Boston shingles and 330 pallets of Vermont shingles.

- a. For the Boston shingles, construct a 95% confidence interval estimate for the mean weight.
- b. For the Vermont shingles, construct a 95% confidence interval estimate for the mean weight.
- c. Do you think the assumption needed to construct the confidence interval estimates in (a) and (b) is valid?
- d. Based on the results of (a) and (b), what conclusions can you reach concerning the mean weight of the Boston and Vermont shingles?

8.68 The manufacturer of Boston and Vermont asphalt shingles provides its customers with a 20-year warranty on most of its products. To determine whether a shingle will last the entire warranty period, accelerated-life testing is conducted at the manufacturing plant. Accelerated-life testing exposes the shingle to the stresses it would be subject to in a lifetime of normal use via a laboratory experiment that takes only a few minutes to conduct. In this test, a shingle is repeatedly scraped with a brush for a short period of time, and the shingle granules removed by the brushing are weighed (in grams). Shingles that experience low amounts of granule loss are expected to last longer in normal use than shingles that experience high amounts of granule loss. In this situation, a shingle should experience no more than 0.8 grams of granule loss if it is expected to last the length of the warranty period. The file **Granule** contains a sample of 170 measurements made on the company's Boston shingles and 140 measurements made on Vermont shingles.

- For the Boston shingles, construct a 95% confidence interval estimate for the mean granule loss.
- For the Vermont shingles, construct a 95% confidence interval estimate for the mean granule loss.
- Do you think the assumption needed to construct the confidence interval estimates in (a) and (b) is valid?
- Based on the results of (a) and (b), what conclusions can you reach concerning the mean granule loss of the Boston and Vermont shingles?

REPORT WRITING EXERCISE

8.69 Referring to the results in Problem 8.66 concerning the width of a steel trough, write a report that summarizes your conclusions.

CASES FOR CHAPTER 8

Managing Ashland MultiComm Services

The marketing department has been considering ways to increase the number of new subscriptions to the *3-For-All* cable/phone/Internet service. Following the suggestion of Assistant Manager Lauren Adler, the department staff designed a survey to help determine various characteristics of households who subscribe to cable television service from Ashland. The survey consists of the following 10 questions:

- Does your household subscribe to telephone service from Ashland?
 (1) Yes (2) No
- Does your household subscribe to Internet service from Ashland?
 (1) Yes (2) No
- What type of cable television service do you have?
 (1) Basic (2) Enhanced
 (If Basic, skip to question 5.)
- How often do you watch the cable television stations that are only available with enhanced service?
 (1) Every day (2) Most days
 (3) Occasionally or never
- How often do you watch premium or on-demand services that require an extra fee?
 (1) Almost every day (2) Several times a week
 (3) Rarely (4) Never
- Which method did you use to obtain your current AMS subscription?

- (1) AMS toll-free phone number
 (2) AMS website
 (3) Direct mail reply card
 (4) Good Tunes & More promotion
 (5) Other
- Would you consider subscribing to the *3-For-All* cable/phone/Internet service for a trial period if a discount were offered?
 (1) Yes (2) No
 (If no, skip to question 9.)
- If purchased separately, cable, Internet, and phone services would currently cost \$24.99 per week. How much would you be willing to pay per week for the *3-For-All* cable/phone/Internet service?
- Does your household use another provider of telephone service?
 (1) Yes (2) No
- AMS may distribute Ashland Gold Cards that would provide discounts at selected Ashland-area restaurants for subscribers who agree to a two-year subscription contract to the *3-For-All* service. Would being eligible to receive a Gold Card cause you to agree to the two-year term?
 (1) Yes (2) No

Of the 500 households selected that subscribe to cable television service from Ashland, 82 households either refused to participate, could not be contacted after repeated attempts,

or had telephone numbers that were not in service. The summary results for the 418 households that were contacted are as follows:

Household Has AMS Telephone Service	Frequency
Yes	83
No	335
Household Has AMS Internet Service	Frequency
Yes	262
No	156
Type of Cable Service	Frequency
Basic	164
Enhanced	254
Watches Enhanced Programming	Frequency
Every day	50
Most days	144
Occasionally or never	60
Watches Premium or On-Demand Services	Frequency
Almost every day	14
Several times a week	35
Almost never	313
Never	56

Method Used to Obtain Current AMS Subscription		Frequency							
Toll-free phone number		230							
AMS website		106							
Direct mail		46							
Good Tunes & More		10							
Other		26							
Would Consider Discounted Trial Offer		Frequency							
Yes		40							
No		378							
Trial Weekly Rate (\$ Willing to Pay (stored in AMS8)									
23.00	20.00	22.75	20.00	20.00	24.50	17.50	22.25	18.00	21.00
18.25	21.00	18.50	20.75	21.25	22.25	22.75	21.75	19.50	20.75
16.75	19.00	22.25	21.00	16.75	19.00	22.25	21.00	19.50	22.75
23.50	19.50	21.75	22.00	24.00	23.25	19.50	20.75	18.25	21.50
Uses Another Phone Service Provider		Frequency							
Yes		354							
No		64							
Gold Card Leads to Two-Year Agreement		Frequency							
Yes		38							
No		380							

Analyze the results of the survey of Ashland households that receive AMS cable television service. Write a report that discusses the marketing implications of the survey results for Ashland MultiComm Services.

Digital Case

Apply your knowledge about confidence interval estimation in this Digital Case, which extends the MyTVLab Digital Case from Chapter 6.

Among its other features, the MyTVLab website allows customers to purchase MyTVLab LifeStyles merchandise online. To handle payment processing, the management of MyTVLab has contracted with the following firms:

- **PayAFriend (PAF)**—This is an online payment system with which customers and businesses such as MyTVLab register in order to exchange payments in a secure and convenient manner, without the need for a credit card.
- **Continental Banking Company (Conbanco)**—This processing services provider allows MyTVLab customers to pay for merchandise using nationally recognized credit cards issued by a financial institution.

To reduce costs, management is considering eliminating one of these two payment systems. However, Lorraine Hildick of the sales department suspects that customers

use the two forms of payment in unequal numbers and that customers display different buying behaviors when using the two forms of payment. Therefore, she would like to first determine the following:

- The proportion of customers using PAF and the proportion of customers using a credit card to pay for their purchases.
- The mean purchase amount when using PAF and the mean purchase amount when using a credit card.

Assist Ms. Hildick by preparing an appropriate analysis. Open **PaymentsSample.pdf**, read Ms. Hildick's comments, and use her random sample of 50 transactions as the basis for your analysis. Summarize your findings to determine whether Ms. Hildick's conjectures about MyTVLab LifeStyle customer purchasing behaviors are correct. If you want the sampling error to be no more than \$3 when estimating the mean purchase amount, is Ms. Hildick's sample large enough to perform a valid analysis?

Sure Value Convenience Stores

You work in the corporate office for a nationwide convenience store franchise that operates nearly 10,000 stores. The per-store daily customer count has been steady, at 900, for some time (i.e., the mean number of customers in a store in one day is 900). To increase the customer count, the franchise is considering cutting coffee prices. The 12-ounce size will now be \$0.59 instead of \$0.99, and the 16-ounce size will be \$0.69 instead of \$1.19. Even with this reduction in price, the franchise will have a 40% gross margin on coffee. To test the new initiative, the franchise has reduced coffee

prices in a sample of 34 stores, where customer counts have been running almost exactly at the national average of 900. After four weeks, the sample stores stabilize at a mean customer count of 974 and a standard deviation of 96. This increase seems like a substantial amount to you, but it also seems like a pretty small sample. Is there some way to get a feel for what the mean per-store count in all the stores will be if you cut coffee prices nationwide? Do you think reducing coffee prices is a good strategy for increasing the mean customer count?

CardioGood Fitness

Return to the CardioGood Fitness case first presented on page 83. Using the data stored in **CardioGood Fitness**:

1. Construct 95% confidence interval estimates to create a customer profile for each CardioGood Fitness treadmill product line.

2. Write a report to be presented to the management of CardioGood Fitness detailing your findings.

More Descriptive Choices Follow-Up

Follow up the More Descriptive Choices, Revisited Using Statistics scenario on page 138 by constructing 95% confidence intervals estimates of the three-year return percentages, five-year return percentages, and ten-year return percentages for the sample of growth and

value funds and for the small, mid-cap, and large market cap funds (stored in **Retirement Funds**). In your analysis, examine differences between the growth and value funds as well as the differences among the small, mid-cap, and large market cap funds.

Clear Mountain State Student Surveys

1. The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. They create and distribute a survey of 14 questions and receive responses from 62 undergraduates (stored in **UndergradSurvey**). For each variable included in the survey, construct a 95% confidence interval estimate for the population characteristic and write a report summarizing your conclusions.

2. The Dean of Students at CMSU has learned about the undergraduate survey and has decided to undertake a similar survey for graduate students at CMSU. She creates and distributes a survey of 14 questions and receives responses from 44 graduate students (stored in **GradSurvey**). For each variable included in the survey, construct a 95% confidence interval estimate for the population characteristic and write a report summarizing your conclusions.

CHAPTER 8 EXCEL GUIDE

EG8.1 CONFIDENCE INTERVAL ESTIMATE for the MEAN (σ KNOWN)

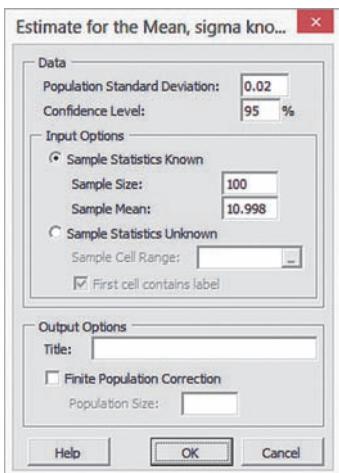
Key Technique Use the **NORM.S.INV(cumulative percentage)** to compute the Z value for one-half of the $(1 - \alpha)$ value and use the **CONFIDENCE(1 - confidence level, population standard deviation, sample size)** function to compute the half-width of a confidence interval.

Example Compute the confidence interval estimate for the mean for the Example 8.1 mean paper length problem on page 277.

PHStat Use Estimate for the Mean, sigma known.

For the example, select **PHStat → Confidence Intervals → Estimate for the Mean, sigma known**. In the procedure's dialog box (shown below):

1. Enter **0.02** as the **Population Standard Deviation**.
2. Enter **95** as the **Confidence Level** percentage.
3. Click **Sample Statistics Known** and enter **100** as the **Sample Size** and **10.998** as the **Sample Mean**.
4. Enter a **Title** and click **OK**.



When using unsummarized data, click **Sample Statistics Unknown** and enter the **Sample Cell Range** in step 3.

In-Depth Excel Use the **COMPUTE worksheet** of the **CIE sigma known workbook** as a template.

The worksheet already contains the data for the example. For other problems, change the **Population Standard Deviation**, **Sample Mean**, **Sample Size**, and **Confidence Level** values in cells B4 through B7. If you use an Excel version older than Excel 2010, use these instructions with the **COMPUTE_OLEDER** worksheet.

EG8.2 CONFIDENCE INTERVAL ESTIMATE for the MEAN (σ UNKNOWN)

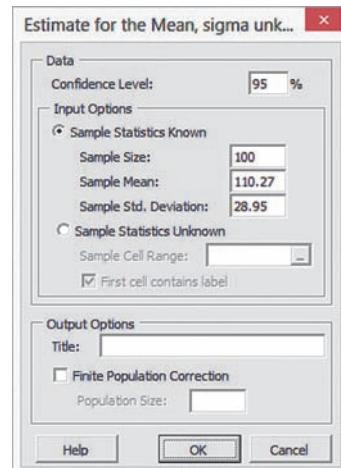
Key Technique Use the **T.INV.2T(1 - confidence level, degrees of freedom)** function to determine the critical value from the *t* distribution.

Example Compute the Figure 8.7 confidence interval estimate for the mean sales invoice amount shown on page 282.

PHStat Use Estimate for the Mean, sigma unknown.

For the example, select **PHStat → Confidence Intervals → Estimate for the Mean, sigma unknown**. In the procedure's dialog box (shown below):

1. Enter **95** as the **Confidence Level** percentage.
2. Click **Sample Statistics Known** and enter **100** as the **Sample Size**, **110.27** as the **Sample Mean**, and **28.95** as the **Sample Std. Deviation**.
3. Enter a **Title** and click **OK**.



When using unsummarized data, click **Sample Statistics Unknown** and enter the **Sample Cell Range** in step 2.

In-Depth Excel Use the **COMPUTE worksheet** of the **CIE sigma unknown workbook** as a template.

The worksheet already contains the data for solving the example. For other problems, change the **Sample Standard Deviation**, **Sample Mean**, **Sample Size**, and **Confidence Level** values in cells B4 through B7. If you use an Excel version older than Excel 2010, use these instructions with the **COMPUTE_OLEDER** worksheet.

EG8.3 CONFIDENCE INTERVAL ESTIMATE for the PROPORTION

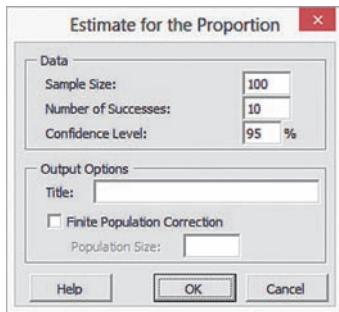
Key Technique Use the **NORM.S.INV((1 – confidence level)/2)** function to compute the Z value.

Example Compute the Figure 8.12 confidence interval estimate for the proportion of in-error sales invoices shown on page 288.

PHStat Use Estimate for the Proportion.

For the example, select **PHStat → Confidence Intervals → Estimate for the Proportion**. In the procedure's dialog box (shown below):

1. Enter **100** as the **Sample Size**.
2. Enter **10** as the **Number of Successes**.
3. Enter **95** as the **Confidence Level** percentage.
4. Enter a **Title** and click **OK**.



In-Depth Excel Use the **COMPUTE worksheet** of the **CIE Proportion workbook** as a template.

The worksheet contains the data for the example. Note that the formula = **SQRT(sample proportion * (1 – sample proportion)/sample size)** computes the standard error of the proportion in cell B11.

To compute confidence interval estimates for other problems, change the **Sample Size**, **Number of Successes**, and **Confidence Level** values in cells B4 through B6. If you use an Excel version older than Excel 2010, use these instructions with the COMPUTE_OLEDR worksheet.

EG8.4 DETERMINING SAMPLE SIZE

Sample Size Determination for the Mean

Key Technique Use the **NORM.S.INV((1 – confidence level)/2)** function to compute the Z value and use the **ROUNDUP(calculated sample size, 0)** function to round up the computed sample size to the next higher integer.

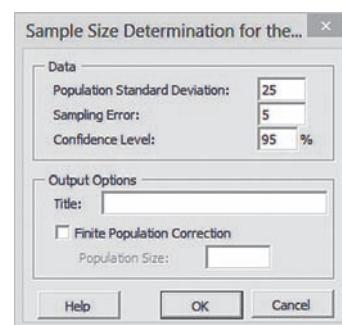
Example Determine the sample size for the mean sales invoice amount example that is shown in Figure 8.13 on page 291.

PHStat Use Determination for the Mean.

For the example, select **PHStat → Sample Size → Determination for the Mean**. In the procedure's dialog box (shown at top right):

1. Enter **25** as the **Population Standard Deviation**.
2. Enter **5** as the **Sampling Error**.

3. Enter **95** as the **Confidence Level** percentage.
4. Enter a **Title** and click **OK**.



In-Depth Excel Use the **COMPUTE worksheet** of the **Sample Size Mean workbook** as a template.

The worksheet already contains the data for the example. For other problems, change the **Population Standard Deviation**, **Sampling Error**, and **Confidence Level** values in cells B4 through B6. If you use an Excel version older than Excel 2010, use these instructions with the COMPUTE_OLEDR worksheet.

Sample Size Determination for the Proportion

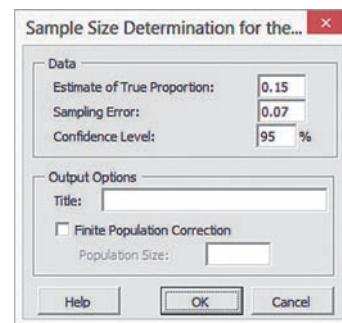
Key Technique Use the **NORM.S.INV** and **ROUNDUP** functions (see previous section) to help determine the sample size needed for estimating the proportion.

Example Determine the sample size for the proportion of in-error sales invoices example that is shown in Figure 8.14 on page 293.

PHStat Use Determination for the Proportion.

For the example, select **PHStat → Sample Size → Determination for the Proportion**. In the procedure's dialog box (shown below):

1. Enter **0.15** as the **Estimate of True Proportion**.
2. Enter **0.07** as the **Sampling Error**.
3. Enter **95** as the **Confidence Level** percentage.
4. Enter a **Title** and click **OK**.



In-Depth Excel Use the **COMPUTE worksheet** of the **Sample Size Proportion workbook** as a template.

The worksheet already contains the data for the example. To compute confidence interval estimates for other problems, change the **Estimate of True Proportion**, **Sampling Error**, and **Confidence Level** in cells B4 through B6. If you use an Excel version older than Excel 2010, use these instructions with the COMPUTE_OLEDR worksheet.

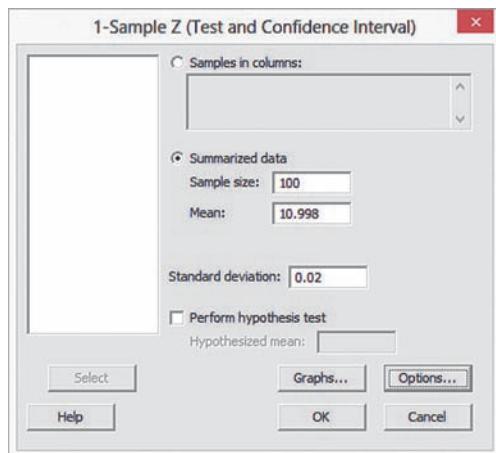
CHAPTER 8 MINITAB GUIDE

MG8.1 CONFIDENCE INTERVAL ESTIMATE for the MEAN (σ KNOWN)

Use 1-Sample Z.

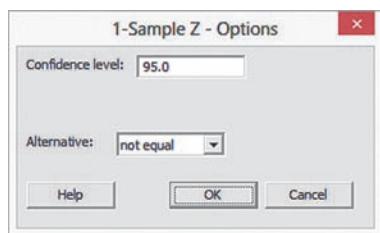
For example, to compute the estimate for the Example 8.1 mean paper length problem on page 277, select **Stat → Basic Statistics → 1-Sample Z**. In the 1-Sample Z (Test and Confidence Interval) dialog box (shown below):

1. Click **Summarized data**.
2. Enter **100** in the **Sample size** box and **110.27** in the **Mean** box.
3. Enter **0.02** in the **Standard deviation** box.
4. Click **Options**.



In the 1-Sample Z - Options dialog box (shown below):

5. Enter **95.0** in the **Confidence level** box.
6. Select **not equal** from the **Alternative** drop-down list.
7. Click **OK**.



8. Back in the original dialog box, click **OK**.

When using unsummarized data, click **Samples in columns** in step 1 and, in step 2, enter the name of the column that contains the data in the **Samples in columns** box.

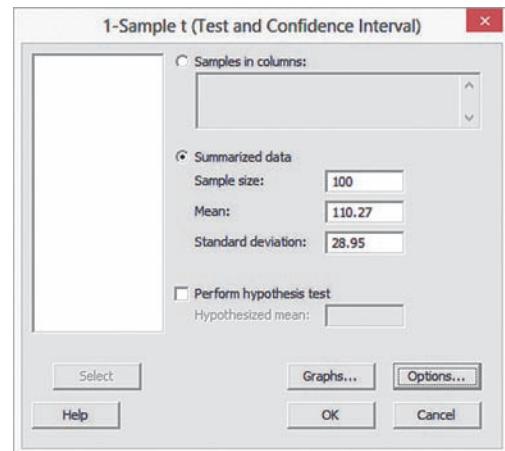
MG8.2 CONFIDENCE INTERVAL ESTIMATE for the MEAN (σ UNKNOWN)

Use 1-Sample t.

For example, to compute the Figure 8.7 estimate for the mean sales invoice amount on page 282, select **Stat → Basic Statistics**

→ **1-Sample t**. In the 1-Sample t (Test and Confidence Interval) dialog box (shown below):

1. Click **Summarized data**.
2. Enter **100** in the **Sample size** box, **110.27** in the **Mean** box, and **28.95** in the **Standard deviation** box.
3. Click **Options**.



In the 1-Sample t - Options dialog box (similar to the 1-Sample Z - Options dialog box shown in left column):

4. Enter **95.0** in the **Confidence level** box.
5. Select **not equal** from the **Alternative** drop-down list.
6. Click **OK**.
7. Back in the original dialog box, click **OK**.

When using unsummarized data, click **Samples in columns** in step 1 and, in step 2, enter the name of the column that contains the data. To create a boxplot of the type shown in Figure 8.9 on page 283, replace step 7 with these steps 7 through 9:

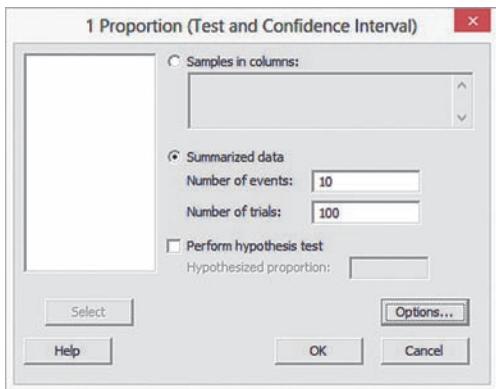
7. Back in the original dialog box, click **Graphs**.
8. In the 1-Sample t - Graphs dialog box, check **Boxplot of data** and then click **OK**.
9. Back in the original dialog box, click **OK**.

MG8.3 CONFIDENCE INTERVAL ESTIMATE for the PROPORTION

Use 1 Proportion.

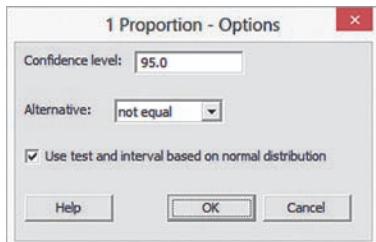
For example, to compute the Figure 8.12 estimate for the proportion of in-error sales invoices on page 288, select **Stat → Basic Statistics → 1 Proportion**. In the 1 Proportion dialog box (shown on page 307):

1. Click **Summarized data**.
2. Enter **10** in the **Number of events** box and **100** in the **Number of trials** box.
3. Click **Options**.



In the 1 Proportion - Options dialog box (shown below):

4. Enter **95.0** in the **Confidence level** box.
5. Select **not equal** from the **Alternative** drop-down list.
6. Check **Use test and interval based on normal distribution**.
7. Click **OK** (to return to the previous dialog box).



8. Back in the original dialog box, click **OK**.

When using unsummarized data, click **Samples in columns** in step 1 and, in step 2, enter the name of the column that contains the data.

MG8.4 DETERMINING SAMPLE SIZE

Minitab version 16 includes Sample Size for Estimation that computes the sample size needed for estimating the mean or the proportion.

To use this new command, select **Stat → Power and Sample Size → Sample Size for Estimation** and in the procedure's dialog box select a parameter from the Parameter drop-down list, complete the entries, and click **OK**. Because this command is not included in Minitab Student 14, the command is not demonstrated or further discussed in this book. (Results using the Minitab 16 command will vary slightly from the Excel results shown in this chapter.)

CHAPTER

9

Fundamentals of Hypothesis Testing: One-Sample Tests

CONTENTS

9.1 Fundamentals of Hypothesis-Testing Methodology

Can You Ever Know the Population Standard Deviation?

9.2 *t* Test of Hypothesis for the Mean (σ Unknown)

9.3 One-Tail Tests

9.4 *Z* Test of Hypothesis for the Proportion

9.5 Potential Hypothesis-Testing Pitfalls and Ethical Issues

9.6 Power of a Test (online)

USING STATISTICS: Significant Testing at Oxford Cereals, Revisited

CHAPTER 9 EXCEL GUIDE

CHAPTER 9 MINITAB GUIDE

OBJECTIVES

To learn the basic principles of hypothesis testing

To learn to use hypothesis testing to test a mean or proportion

To know the assumptions of each hypothesis-testing procedure, how to evaluate them, and the consequences if they are seriously violated

To be aware of the pitfalls and ethical issues involved in hypothesis testing

To learn to avoid the pitfalls involved in hypothesis testing

USING STATISTICS

Significant Testing at Oxford Cereals

As in Chapter 7, you again find yourself as plant operations manager for Oxford Cereals. Among other responsibilities, you are responsible for monitoring the amount in each cereal box filled. Company specifications require a mean weight of 368 grams per box. You must adjust the cereal-filling process when the mean fill weight in the population of boxes differs from 368 grams. Adjusting the process requires shutting down the cereal production line temporarily, so you do not want to make unnecessary adjustments.

What decision-making method can you use to decide if the cereal-filling process needs to be adjusted? You decide to begin by selecting a random sample of 25 cereal boxes and weighing each box. From the weights collected, you compute a sample mean. How could that sample mean be used to help decide whether adjustment is necessary?



Peter Close/Shutterstock

In Chapter 7, you learned methods to determine whether the value of a sample mean is consistent with a known population mean. In this Oxford Cereals scenario, you seek to use a sample mean to validate a claim about the population mean, a somewhat different problem. For this type of situation, you use the inferential method known as **hypothesis testing**. Hypothesis testing requires that you state a claim unambiguously. In this scenario, the claim is that the population mean is 368 grams. You examine a sample statistic to see if it better supports the stated claim, called the *null hypothesis*, or the mutually exclusive alternative hypothesis (for this scenario, that the population mean is not 368 grams).

In this chapter, you will learn several applications of hypothesis testing. You will learn how to make inferences about a population parameter by *analyzing differences* between the results observed, the sample statistic, and the results you would expect to get if an underlying hypothesis were actually true. For the Oxford Cereals scenario, hypothesis testing allows you to infer one of the following:

- The mean weight of the cereal boxes in the sample is a value consistent with what you would expect if the mean of the entire population of cereal boxes were 368 grams.
- The population mean is not equal to 368 grams because the sample mean is significantly different from 368 grams.

9.1 Fundamentals of Hypothesis-Testing Methodology

Hypothesis testing typically begins with a theory, a claim, or an assertion about a particular parameter of a population. For example, your initial hypothesis in the cereal example is that the process is working properly, so the mean fill is 368 grams, and no corrective action is needed.

The Null and Alternative Hypotheses

The hypothesis that the population parameter is equal to the company specification is referred to as the null hypothesis. A **null hypothesis** is often one of status quo and is identified by the symbol H_0 . Here the null hypothesis is that the filling process is working properly, and therefore the mean fill is the 368-gram specification provided by Oxford Cereals. This is stated as

$$H_0 : \mu = 368$$

Student Tip

Remember, hypothesis testing reaches conclusions about parameters, not statistics.

Even though information is available only from the sample, the null hypothesis is stated in terms of the population parameter because your focus is on the population of all cereal boxes. You use the sample statistic to make inferences about the entire filling process. One inference may be that the results observed from the sample data indicate that the null hypothesis is false. If the null hypothesis is considered false, something else must be true.

Whenever a null hypothesis is specified, an alternative hypothesis is also specified, and it must be true if the null hypothesis is false. The **alternative hypothesis**, H_1 , is the opposite of the null hypothesis, H_0 . This is stated in the cereal example as

$$H_1 : \mu \neq 368$$

The alternative hypothesis represents the conclusion reached by rejecting the null hypothesis. In many research situations, the alternative hypothesis serves as the hypothesis that is the focus of the research being conducted. The null hypothesis is rejected when there is sufficient evidence from the sample data that the null hypothesis is false. In the cereal example, if the weights of the sampled boxes are sufficiently above or below the expected 368-gram mean specified by Oxford Cereals, you reject the null hypothesis in favor of the alternative hypothesis that the mean fill is different from 368 grams. You stop production and take whatever action is necessary to correct the problem. If the null hypothesis is not rejected, you should continue to believe that the process is working correctly and that therefore no corrective action is necessary. In this second circumstance, you have not proven that the process is working correctly.

Rather, you have failed to prove that it is working incorrectly, and therefore you continue your belief (although unproven) in the null hypothesis.

In hypothesis testing, you reject the null hypothesis when the sample evidence suggests that it is far more likely that the alternative hypothesis is true. However, failure to reject the null hypothesis is not proof that it is true. You can never prove that the null hypothesis is correct because the decision is based only on the sample information, not on the entire population. Therefore, if you fail to reject the null hypothesis, you can only conclude that there is insufficient evidence to warrant its rejection. The following key points summarize the null and alternative hypotheses:

- The null hypothesis, H_0 , represents the current belief in a situation.
- The alternative hypothesis, H_1 , is the opposite of the null hypothesis and represents a research claim or specific inference you would like to prove.
- If you reject the null hypothesis, you have statistical proof that the alternative hypothesis is correct.
- If you do not reject the null hypothesis, you have failed to prove the alternative hypothesis. The failure to prove the alternative hypothesis, however, does not mean that you have proven the null hypothesis.
- The null hypothesis, H_0 , always refers to a specified value of the population parameter (such as μ), not a sample statistic (such as \bar{X}).
- The statement of the null hypothesis always contains an equal sign regarding the specified value of the population parameter (e.g., $H_0 : \mu = 368$ grams).
- The statement of the alternative hypothesis never contains an equal sign regarding the specified value of the population parameter (e.g., $H_1 : \mu \neq 368$ grams).

EXAMPLE 9.1

The Null and Alternative Hypotheses

You are the manager of a fast-food restaurant. You want to determine whether the waiting time to place an order has changed in the past month from its previous population mean value of 4.5 minutes. State the null and alternative hypotheses.

SOLUTION The null hypothesis is that the population mean has not changed from its previous value of 4.5 minutes. This is stated as

$$H_0 : \mu = 4.5$$

The alternative hypothesis is the opposite of the null hypothesis. Because the null hypothesis is that the population mean is 4.5 minutes, the alternative hypothesis is that the population mean is not 4.5 minutes. This is stated as

$$H_1 : \mu \neq 4.5$$

The Critical Value of the Test Statistic

Hypothesis testing uses sample data to determine how likely it is that the null hypothesis is true. In the Oxford Cereal Company scenario, the null hypothesis is that the mean amount of cereal per box in the entire filling process is 368 grams (the population parameter specified by the company). You select a sample of boxes from the filling process, weigh each box, and compute the sample mean \bar{X} . This sample statistic is an estimate of the corresponding parameter, the population mean, μ . Even if the null hypothesis is true, the sample statistic \bar{X} is likely to differ from the value of the parameter (the population mean, μ) because of variation due to sampling.

You do expect the sample statistic to be close to the population parameter if the null hypothesis is true. If the sample statistic is close to the population parameter, you have insufficient evidence to reject the null hypothesis. For example, if the sample mean is 367.9 grams, you might conclude that the population mean has not changed (i.e., $\mu = 368$) because a sample mean of 367.9 grams is very close to the hypothesized value of 368 grams. Intuitively, you think that it is likely that you could get a sample mean of 367.9 grams from a population whose mean is 368.

However, if there is a large difference between the value of the sample statistic and the hypothesized value of the population parameter, you might conclude that the null hypothesis is false. For example, if the sample mean is 320 grams, you might conclude that the population mean is not 368 grams (i.e., $\mu \neq 368$) because the sample mean is very far from the hypothesized value of 368 grams. In such a case, you might conclude that it is very unlikely to get a sample mean of 320 grams if the population mean is really 368 grams. Therefore, it is more logical to conclude that the population mean is not equal to 368 grams. Here you reject the null hypothesis.

However, the decision-making process is not always so clear-cut. Determining what is “very close” and what is “very different” is arbitrary without clear definitions. Hypothesis-testing methodology provides clear definitions for evaluating differences. Furthermore, it enables you to quantify the decision-making process by computing the probability of getting a certain sample result if the null hypothesis is true. You calculate this probability by determining the sampling distribution for the sample statistic of interest (e.g., the sample mean) and then computing the particular **test statistic** based on the given sample result. Because the sampling distribution for the test statistic often follows a well-known statistical distribution, such as the standardized normal distribution or *t* distribution, you can use these distributions to help determine whether the null hypothesis is true.

Student Tip

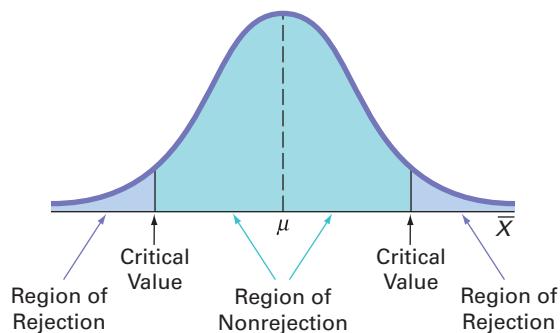
Every test statistic follows a specific sampling distribution.

Regions of Rejection and Nonrejection

The sampling distribution of the test statistic is divided into two regions, a **region of rejection** (sometimes called the critical region) and a **region of nonrejection** (see Figure 9.1). If the test statistic falls into the region of nonrejection, you do not reject the null hypothesis. In the Oxford Cereals scenario, you conclude that there is insufficient evidence that the population mean fill is different from 368 grams. If the test statistic falls into the rejection region, you reject the null hypothesis. In this case, you conclude that the population mean is not 368 grams.

FIGURE 9.1

Regions of rejection and nonrejection in hypothesis testing



The region of rejection consists of the values of the test statistic that are unlikely to occur if the null hypothesis is true. These values are much more likely to occur if the null hypothesis is false. Therefore, if a value of the test statistic falls into this rejection region, you reject the null hypothesis because that value is unlikely if the null hypothesis is true.

To make a decision concerning the null hypothesis, you first determine the **critical value** of the test statistic. The critical value divides the nonrejection region from the rejection region. Determining the critical value depends on the size of the rejection region. The size of the rejection region is directly related to the risks involved in using only sample evidence to make decisions about a population parameter.

Risks in Decision Making Using Hypothesis Testing

Using hypothesis testing involves the risk of reaching an incorrect conclusion. You might wrongly reject a true null hypothesis, H_0 , or, conversely, you might wrongly *not* reject a false null hypothesis, H_0 . These types of risk are called Type I and Type II errors.

TYPE I AND TYPE II ERRORS

A **Type I error** occurs if you reject the null hypothesis, H_0 , when it is true and should not be rejected. A Type I error is a “false alarm.” The probability of a Type I error occurring is α .

A **Type II error** occurs if you do not reject the null hypothesis, H_0 , when it is false and should be rejected. A Type II error represents a “missed opportunity” to take some corrective action. The probability of a Type II error occurring is β .

In the Oxford Cereals scenario, you would make a Type I error if you concluded that the population mean fill is *not* 368 grams when it *is* 368 grams. This error causes you to needlessly adjust the filling process (the “false alarm”) even though the process is working properly. In the same scenario, you would make a Type II error if you concluded that the population mean fill *is* 368 grams when it *is not* 368 grams. In this case, you would allow the process to continue without adjustment, even though an adjustment is needed (the “missed opportunity”).

Traditionally, you control the Type I error by determining the risk level, α (the lowercase Greek letter *alpha*), that you are willing to have of rejecting the null hypothesis when it is true. This risk, or probability, of committing a Type I error is called the *level of significance* (α). Because you specify the level of significance before you perform the hypothesis test, you directly control the risk of committing a Type I error. Traditionally, you select a level of 0.01, 0.05, or 0.10. The choice of a particular risk level for making a Type I error depends on the cost of making a Type I error. After you specify the value for α , you can then determine the critical values that divide the rejection and nonrejection regions. You know the size of the rejection region because α is the probability of rejection when the null hypothesis is true. From this, you can then determine the critical value or values that divide the rejection and nonrejection regions.

The probability of committing a Type II error is called the β *risk*. Unlike with a Type I error, which you control through the selection of α , the probability of making a Type II error depends on the difference between the hypothesized and actual values of the population parameter. Because large differences are easier to find than small ones, if the difference between the hypothesized and actual values of the population parameter is large, β is small. For example, if the population mean is 330 grams, there is a small chance (β) that you will conclude that the mean has not changed from 368 grams. However, if the difference between the hypothesized and actual values of the parameter is small, β is large. For example, if the population mean is actually 367 grams, there is a large chance (β) that you will conclude that the mean is still 368 grams.

PROBABILITY OF TYPE I AND TYPE II ERRORS

The **level of significance** (α) of a statistical test is the probability of committing a Type I error.

The **β risk** is the probability of committing a Type II error.

The complement of the probability of a Type I error, $(1 - \alpha)$, is called the *confidence coefficient*. The confidence coefficient is the probability that you will not reject the null hypothesis, H_0 , when it is true and should not be rejected. In the Oxford Cereals scenario, the confidence coefficient measures the probability of concluding that the population mean fill is 368 grams when it is actually 368 grams.

The complement of the probability of a Type II error, $(1 - \beta)$, is called the *power of a statistical test*. The power of a statistical test is the probability that you will reject the null hypothesis when it is false and should be rejected. In the Oxford Cereals scenario, the power of the test is the probability that you will correctly conclude that the mean fill amount is not 368 grams when it actually is not 368 grams.

COMPLEMENTS OF TYPE I AND TYPE II ERRORS

The **confidence coefficient**, $(1 - \alpha)$, is the probability that you will not reject the null hypothesis, H_0 , when it is true and should not be rejected.

The **power of a statistical test**, $(1 - \beta)$, is the probability that you will reject the null hypothesis when it is false and should be rejected.

Table 9.1 illustrates the results of the two possible decisions (do not reject H_0 or reject H_0) that you can make in any hypothesis test. You can make a correct decision or make one of two types of errors.

TABLE 9.1

Hypothesis Testing
and Decision Making

STATISTICAL DECISION	ACTUAL SITUATION	
	H_0 True	H_0 False
Do not reject H_0	Correct decision $\text{Confidence} = (1 - \alpha)$	Type II error $P(\text{Type II error}) = \beta$
Reject H_0	Type I error $P(\text{Type I error}) = \alpha$	Correct decision Power = $(1 - \beta)$

One way to reduce the probability of making a Type II error is by increasing the sample size. Large samples generally permit you to detect even very small differences between the hypothesized values and the actual population parameters. For a given level of α , increasing the sample size decreases β and therefore increases the power of the statistical test to detect that the null hypothesis, H_0 , is false.

However, there is always a limit to your resources, and this affects the decision of how large a sample you can select. For any given sample size, you must consider the trade-offs between the two possible types of errors. Because you can directly control the risk of a Type I error, you can reduce this risk by selecting a smaller value for α . For example, if the negative consequences associated with making a Type I error are substantial, you could select $\alpha = 0.01$ instead of 0.05. However, when you decrease α , you increase β , so reducing the risk of a Type I error results in an increased risk of a Type II error. However, to reduce β , you could select a larger value for α . Therefore, if it is important to try to avoid a Type II error, you can select α of 0.05 or 0.10 instead of 0.01.

In the Oxford Cereals scenario, the risk of a Type I error occurring involves concluding that the mean fill amount has changed from the hypothesized 368 grams when it actually has not changed. The risk of a Type II error occurring involves concluding that the mean fill amount has not changed from the hypothesized 368 grams when it actually has changed. The choice of reasonable values for α and β depends on the costs inherent in each type of error. For example, if it is very costly to change the cereal-filling process, you would want to be very confident that a change is needed before making any changes. In this case, the risk of a Type I error occurring is more important, and you would choose a small α . However, if you want to be very certain of detecting changes from a mean of 368 grams, the risk of a Type II error occurring is more important, and you would choose a higher level of α .

Now that you have been introduced to hypothesis testing, recall that in the Oxford Cereals scenario on page 308, the business problem facing Oxford Cereals is to determine if the mean fill weight in the population of boxes in the cereal-filling process differs from 368 grams. To make this determination, you select a random sample of 25 boxes, weigh each box, compute the sample mean, \bar{X} , and then evaluate the difference between this sample statistic and the hypothesized population parameter by comparing the sample mean weight (in grams) to the expected population mean of 368 grams specified by the company. The null and alternative hypotheses are:

$$H_0 : \mu = 368$$

$$H_1 : \mu \neq 368$$

Z Test for the Mean (σ Known)

When the standard deviation, σ , is known (which rarely occurs), you use the **Z test for the mean** if the population is normally distributed. If the population is not normally distributed, you can still use the Z test if the sample size is large enough for the Central Limit Theorem to take effect (see Section 7.2). Equation (9.1) defines the Z_{STAT} test statistic for determining the difference between the sample mean, \bar{X} , and the population mean, μ , when the standard deviation, σ , is known.

Z TEST FOR THE MEAN (σ KNOWN)

$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad (9.1)$$

In Equation (9.1), the numerator measures the difference between the observed sample mean, \bar{X} , and the hypothesized mean, μ . The denominator is the standard error of the mean, so Z_{STAT} represents the difference between \bar{X} and μ in standard error units.

Hypothesis Testing Using the Critical Value Approach

The critical value approach compares the value of the computed Z_{STAT} test statistic from Equation (9.1) to critical values that divide the normal distribution into regions of rejection and nonrejection. The critical values are expressed as standardized Z values that are determined by the level of significance.

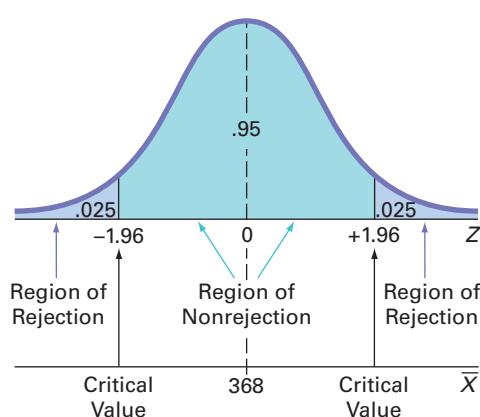
For example, if you use a level of significance of 0.05, the size of the rejection region is 0.05. Because the null hypothesis contains an equal sign and the alternative hypothesis contains a not equal sign, you have a **two-tail test** in which the rejection region is divided into the two tails of the distribution, with two equal parts of 0.025 in each tail. For this two-tail test, a rejection region of 0.025 in each tail of the normal distribution results in a cumulative area of 0.025 below the lower critical value and a cumulative area of 0.975 ($1 - 0.025$) below the upper critical value (which leaves an area of 0.025 in the upper tail). According to the cumulative standardized normal distribution table (Table E.2), the critical values that divide the rejection and nonrejection regions are -1.96 and $+1.96$. Figure 9.2 illustrates that if the mean is actually 368 grams, as H_0 claims, the values of the Z_{STAT} test statistic have a standardized normal distribution centered at $Z = 0$ (which corresponds to an \bar{X} value of 368 grams). Values of Z_{STAT} greater than $+1.96$ and less than -1.96 indicate that \bar{X} is sufficiently different from the hypothesized $\mu = 368$ that it is unlikely that such an \bar{X} value would occur if H_0 were true.

Student Tip

Remember, first you determine the level of significance. This enables you to then determine the critical value. A different level of significance leads to a different critical value.

FIGURE 9.2

Testing a hypothesis about the mean (σ known) at the 0.05 level of significance



Student Tip

In a two-tail test, there is a rejection region in each tail of the distribution.

Therefore, the decision rule is

Reject H_0 if $Z_{STAT} > +1.96$
or if $Z_{STAT} < -1.96$;
otherwise, do not reject H_0 .

Suppose that the sample of 25 cereal boxes indicates a sample mean, \bar{X} , of 372.5 grams, and the population standard deviation, σ , is 15 grams. Using Equation (9.1) on page 314,

$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{372.5 - 368}{\frac{15}{\sqrt{25}}} = +1.50$$

Because $Z_{STAT} = +1.50$ is greater than -1.96 and less than $+1.96$, you do not reject H_0 (see Figure 9.3).

You continue to believe that the mean fill amount is 368 grams. To take into account the possibility of a Type II error, you state the conclusion as “there is insufficient evidence that the mean fill is different from 368 grams.”

Student Tip

Remember, the decision always concerns H_0 . Either you reject H_0 or you do not reject H_0 .

FIGURE 9.3

Testing a hypothesis about the mean cereal weight (σ known) at the 0.05 level of significance

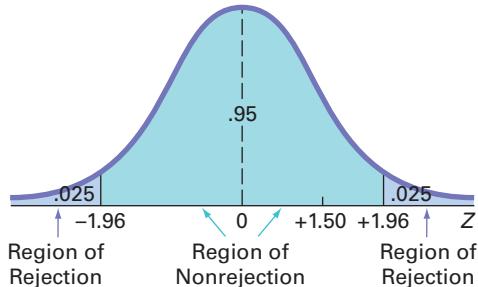


Exhibit 9.1 summarizes the critical value approach to hypothesis testing. Steps 1 and 2 are part of the Define task, step 5 combines the Collect and Organize tasks, and steps 3, 4, and 6 involve the Visualize and Analyze tasks of the DCOVA framework first introduced on page 2. Examples 9.2 and 9.3 apply the critical value approach to hypothesis testing to Oxford Cereals and to a fast-food restaurant.

EXHIBIT 9.1

The Critical Value Approach to Hypothesis Testing

- Step 1** State the null hypothesis, H_0 , and the alternative hypothesis, H_1 .
- Step 2** Choose the level of significance, α , and the sample size, n . The level of significance is based on the relative importance of the risks of committing Type I and Type II errors in the problem.
- Step 3** Determine the appropriate test statistic and sampling distribution.
- Step 4** Determine the critical values that divide the rejection and nonrejection regions.
- Step 5** Collect the sample data, organize the results, and compute the value of the test statistic.
- Step 6** Make the statistical decision, determine whether the assumptions are valid, and state the managerial conclusion in the context of the theory, claim, or assertion being tested. If the test statistic falls into the nonrejection region, you do not reject the null hypothesis. If the test statistic falls into the rejection region, you reject the null hypothesis.

EXAMPLE 9.2

Applying the Critical Value Approach to Hypothesis Testing at Oxford Cereals

State the critical value approach to hypothesis testing at Oxford Cereals.

SOLUTION

- Step 1** State the null and alternative hypotheses. The null hypothesis, H_0 , is always stated as a mathematical expression, using population parameters. In testing whether the mean fill is 368 grams, the null hypothesis states that μ equals 368. The alternative hypothesis, H_1 , is also stated as a mathematical expression, using population parameters. Therefore, the alternative hypothesis states that μ is not equal to 368 grams.
- Step 2** Choose the level of significance and the sample size. You choose the level of significance, α , according to the relative importance of the risks of committing Type I and Type II errors in the problem. The smaller the value of α , the less risk there is of making a Type I error. In this example, making a Type I error means that you conclude that the population mean is not 368 grams when it is 368 grams. Thus, you will take corrective action on the filling process even though the process is working properly. Here, $\alpha = 0.05$ is selected. The sample size, n , is 25.
- Step 3** Select the appropriate test statistic. Because σ is known from information about the filling process, you use the normal distribution and the Z_{STAT} test statistic.
- Step 4** Determine the rejection region. Critical values for the appropriate test statistic are selected so that the rejection region contains a total area of α when H_0 is true and the nonrejection region contains a total area of $1 - \alpha$ when H_0 is true. Because $\alpha = 0.05$ in the cereal example, the critical values of the Z_{STAT} test statistic are -1.96 and $+1.96$. The rejection region is therefore $Z_{STAT} < -1.96$ or $Z_{STAT} > +1.96$. The nonrejection region is $-1.96 \leq Z_{STAT} \leq +1.96$.
- Step 5** Collect the sample data and compute the value of the test statistic. In the cereal example, $\bar{X} = 372.5$, and the value of the test statistic is $Z_{STAT} = +1.50$.
- Step 6** State the statistical decision and the managerial conclusion. First, determine whether the test statistic has fallen into the rejection region or the nonrejection region. For the cereal example, $Z_{STAT} = +1.50$ is in the region of nonrejection because $-1.96 \leq Z_{STAT} = +1.50 \leq +1.96$. Because the test statistic falls into the nonrejection region, the statistical decision is to not reject the null hypothesis, H_0 . The managerial conclusion is that insufficient evidence exists to prove that the mean fill is different from 368 grams. No corrective action on the filling process is needed.

EXAMPLE 9.3

Testing and Rejecting a Null Hypothesis

You are the manager of a fast-food restaurant. The business problem is to determine whether the population mean waiting time to place an order has changed in the past month from its previous population mean value of 4.5 minutes. From past experience, you can assume that the population is normally distributed, with a population standard deviation of 1.2 minutes. You select a sample of 25 orders during a one-hour period. The sample mean is 5.1 minutes. Use the six-step approach listed in Exhibit 9.1 on page 315 to determine whether there is evidence at the 0.05 level of significance that the population mean waiting time to place an order has changed in the past month from its previous population mean value of 4.5 minutes.

SOLUTION

- Step 1** The null hypothesis is that the population mean has not changed from its previous value of 4.5 minutes:

$$H_0 : \mu = 4.5$$

The alternative hypothesis is the opposite of the null hypothesis. Because the null hypothesis is that the population mean is 4.5 minutes, the alternative hypothesis is that the population mean is not 4.5 minutes:

$$H_1 : \mu \neq 4.5$$

- Step 2** You have selected a sample of $n = 25$. The level of significance is 0.05 (i.e., $\alpha = 0.05$).
- Step 3** Because σ is assumed to be known, you use the normal distribution and the Z_{STAT} test statistic.
- Step 4** Because $\alpha = 0.05$, the critical values of the Z_{STAT} test statistic are -1.96 and $+1.96$. The rejection region is $Z_{STAT} < -1.96$ or $Z_{STAT} > +1.96$. The nonrejection region is $-1.96 \leq Z_{STAT} \leq +1.96$
- Step 5** You collect the sample data and compute $\bar{X} = 5.1$. Using Equation (9.1) on page 314, you compute the test statistic:

$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{5.1 - 4.5}{\frac{1.2}{\sqrt{25}}} = +2.50$$

- Step 6** Because $Z_{STAT} = +2.50 > +1.96$, you reject the null hypothesis. You conclude that there is evidence that the population mean waiting time to place an order has changed from its previous value of 4.5 minutes. The mean waiting time for customers is longer now than it was last month. As the manager, you would now want to determine how waiting time could be reduced to improve service.

Hypothesis Testing Using the *p*-Value Approach

The ***p*-value** is the probability of getting a test statistic equal to or more extreme than the sample result, given that the null hypothesis, H_0 , is true. The *p*-value is also known as the *observed level of significance*. Using the *p*-value to determine rejection and nonrejection is another approach to hypothesis testing.

The decision rules for rejecting H_0 in the *p*-value approach are

- If the *p*-value is greater than or equal to α , do not reject the null hypothesis.
- If the *p*-value is less than α , reject the null hypothesis.

Many people confuse these rules, mistakenly believing that a high *p*-value is reason for rejection. You can avoid this confusion by remembering the following:

If the *p*-value is low, then H_0 must go.

Student Tip

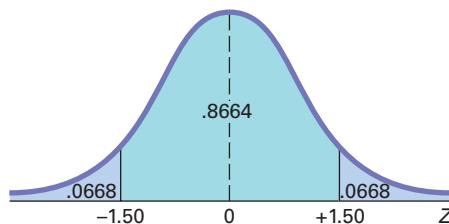
A small (or low) *p*-value indicates a small probability that H_0 is true. A big or large *p*-value indicates a large probability that H_0 is true.

To understand the *p*-value approach, consider the Oxford Cereals scenario. You tested whether the mean fill was equal to 368 grams. The test statistic resulted in a Z_{STAT} value of $+1.50$ and you did not reject the null hypothesis because $+1.50$ was less than the upper critical value of $+1.96$ and greater than the lower critical value of -1.96 .

To use the *p*-value approach for the *two-tail test*, you find the probability that the test statistic Z_{STAT} is equal to or *more extreme than* 1.50 standard error units from the center of a standardized normal distribution. In other words, you need to compute the probability that the Z_{STAT} value is greater than $+1.50$ along with the probability that the Z_{STAT} value is less than -1.50 . Table E.2 shows that the probability of a Z_{STAT} value below -1.50 is 0.0668. The probability of a value below $+1.50$ is 0.9332, and the probability of a value above $+1.50$ is $1 - 0.9332 = 0.0668$. Therefore, the *p*-value for this two-tail test is $0.0668 + 0.0668 = 0.1336$ (see Figure 9.4). Thus, the probability of a test statistic equal to or more extreme than the sample result is 0.1336. Because 0.1336 is greater than $\alpha = 0.05$, you do not reject the null hypothesis.

FIGURE 9.4

Finding a *p*-value for a two-tail test



In this example, the observed sample mean is 372.5 grams, 4.5 grams above the hypothesized value, and the *p*-value is 0.1336. Thus, if the population mean is 368 grams, there is a 13.36% chance that the sample mean differs from 368 grams by at least 4.5 grams (i.e., is ≥ 372.5 grams or ≤ 363.5 grams). Therefore, even though 372.5 grams is above the hypothesized value of 368 grams, a result as extreme as or more extreme than 372.5 grams is not highly unlikely when the population mean is 368 grams.

Unless you are dealing with a test statistic that follows the normal distribution, you will only be able to approximate the *p*-value from the tables of the distribution. However, Excel and Minitab can compute the *p*-value for any hypothesis test, and this allows you to substitute the *p*-value approach for the critical value approach when you conduct hypothesis testing.

Figure 9.5 displays the Excel and Minitab results for the cereal-filling example discussed beginning on page 314.

FIGURE 9.5

Excel and Minitab results for the *Z* test for the mean (σ known) for the cereal-filling example

A	B	One-Sample Z
1 Z Test for the Mean		Test of $\mu = 368$ vs not = 368
2		The assumed standard deviation = 15
3 Data		
4 Null Hypothesis $\mu =$	368	N 372.50
5 Level of Significance	0.05	Mean 3.00
6 Population Standard Deviation	15	95% CI (366.62, 378.38)
7 Sample Size	25	Z 1.50
8 Sample Mean	372.5	P 0.134
9		
10 Intermediate Calculations		
11 Standard Error of the Mean	3	=B6/SQRT(B7)
12 Z Test Statistic	1.5	=(B8 - B4)/B11
13		
14 Two-Tail Test		
15 Lower Critical Value	-1.9600	=NORM.S.INV(B5/2)
16 Upper Critical Value	1.9600	=NORM.S.INV(1 - B5/2)
17 <i>p</i> -Value	0.1336	=2 * (1 - NORM.S.DIST(ABS(B12), TRUE))
18 Do not reject the null hypothesis		=IF(B17 < B5, "Reject the null hypothesis", "Do not reject the null hypothesis")

Exhibit 9.2 summarizes the *p*-value approach to hypothesis testing. Example 9.4 applies the *p*-value approach to the fast-food restaurant example.

EXHIBIT 9.2

The *p*-Value Approach to Hypothesis Testing

- Step 1** State the null hypothesis, H_0 , and the alternative hypothesis, H_1 .
- Step 2** Choose the level of significance, α , and the sample size, n . The level of significance is based on the relative importance of the risks of committing Type I and Type II errors in the problem.
- Step 3** Determine the appropriate test statistic and the sampling distribution.
- Step 4** Collect the sample data, compute the value of the test statistic, and compute the *p*-value.
- Step 5** Make the statistical decision and state the managerial conclusion in the context of the theory, claim, or assertion being tested. If the *p*-value is greater than or equal to α , do not reject the null hypothesis. If the *p*-value is less than α , reject the null hypothesis.

EXAMPLE 9.4

Testing and Rejecting a Null Hypothesis Using the *p*-Value Approach

You are the manager of a fast-food restaurant. The business problem is to determine whether the population mean waiting time to place an order has changed in the past month from its previous value of 4.5 minutes. From past experience, you can assume that the population standard deviation is 1.2 minutes and the population waiting time is normally distributed. You select a sample of 25 orders during a one-hour period. The sample mean is 5.1 minutes. Use the five-step *p*-value approach of Exhibit 9.2 to determine whether there is evidence that the population mean waiting time to place an order has changed in the past month from its previous population mean value of 4.5 minutes.

SOLUTION

Step 1 The null hypothesis is that the population mean has not changed from its previous value of 4.5 minutes:

$$H_0 : \mu = 4.5$$

The alternative hypothesis is the opposite of the null hypothesis. Because the null hypothesis is that the population mean is 4.5 minutes, the alternative hypothesis is that the population mean is not 4.5 minutes:

$$H_1 : \mu \neq 4.5$$

Step 2 You have selected a sample of $n = 25$ and you have chosen a 0.05 level of significance (i.e., $\alpha = 0.05$).

Step 3 Select the appropriate test statistic. Because σ is assumed known, you use the normal distribution and the Z_{STAT} test statistic.

Step 4 You collect the sample data and compute $\bar{X} = 5.1$. Using Equation (9.1) on page 314, you compute the test statistic as follows:

$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{5.1 - 4.5}{\frac{1.2}{\sqrt{25}}} = +2.50$$

To find the probability of getting a Z_{STAT} test statistic that is equal to or more extreme than 2.50 standard error units from the center of a standardized normal distribution, you compute the probability of a Z_{STAT} value greater than +2.50 along with the probability of a Z_{STAT} value less than -2.50. From Table E.2, the probability of a Z_{STAT} value below -2.50 is 0.0062. The probability of a value below +2.50 is 0.9938. Therefore, the probability of a value above +2.50 is $1 - 0.9938 = 0.0062$. Thus, the *p*-value for this two-tail test is $0.0062 + 0.0062 = 0.0124$.

Step 5 Because the *p*-value = $0.0124 < \alpha = 0.05$, you reject the null hypothesis. You conclude that there is evidence that the population mean waiting time to place an order has changed from its previous population mean value of 4.5 minutes. The mean waiting time for customers is longer now than it was last month.

A Connection Between Confidence Interval Estimation and Hypothesis Testing

This chapter and Chapter 8 discuss confidence interval estimation and hypothesis testing, the two major elements of statistical inference. Although confidence interval estimation and hypothesis testing share the same conceptual foundation, they are used for different purposes. In Chapter 8, confidence intervals estimated parameters. In this chapter, hypothesis testing makes decisions about specified values of population parameters. Hypothesis tests are used when trying to determine whether a parameter is less than, more than, or not equal to a specified value. Proper interpretation of a confidence interval, however, can also indicate whether a parameter is less than, more than, or not equal to a specified value. For example, in

this section, you tested whether the population mean fill amount was different from 368 grams by using Equation (9.1) on page 314:

$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Instead of testing the null hypothesis that $\mu = 368$ grams, you can reach the same conclusion by constructing a confidence interval estimate of μ . If the hypothesized value of $\mu = 368$ is contained within the interval, you do not reject the null hypothesis because 368 would not be considered an unusual value. However, if the hypothesized value does not fall into the interval, you reject the null hypothesis because $\mu = 368$ grams is then considered an unusual value. Using Equation (8.1) on page 276 and the following results:

$$n = 25, \bar{X} = 372.5 \text{ grams}, \sigma = 15 \text{ grams}$$

for a confidence level of 95% (i.e., $\alpha = 0.05$),

$$\begin{aligned} \bar{X} &\pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ 372.5 &\pm (1.96) \frac{15}{\sqrt{25}} \\ 372.5 &\pm 5.88 \end{aligned}$$

so that

$$366.62 \leq \mu \leq 378.38$$

Because the interval includes the hypothesized value of 368 grams, you do not reject the null hypothesis. There is insufficient evidence that the mean fill amount for the entire filling process is not 368 grams. You reached the same decision by using a two-tail hypothesis test.

Can You Ever Know the Population Standard Deviation?

The end of Section 8.1 on page 278 discussed how learning a confidence interval estimation method that required knowing σ , the population standard deviation, served as an effective introduction to the concept of a confidence interval. That section then revealed that you would be unlikely to use that procedure for most practical applications for several reasons.

Likewise, for most practical applications, you are unlikely to use a hypothesis-testing method that requires knowing σ . If you knew the population standard deviation, you would also know the population mean and would not need to form a hypothesis about the mean and

then test that hypothesis. So why study a hypothesis testing of the mean, which requires that σ is known? Using such a test makes it much easier to explain the fundamentals of hypothesis testing. With a known population standard deviation, you can use the normal distribution and compute p -values using the tables of the normal distribution.

Because it is important that you understand the concept of hypothesis testing when reading the rest of this book, review this section carefully—even if you anticipate never having a practical reason to use the test represented in Equation (9.1).

Problems for Section 9.1

LEARNING THE BASICS

9.1 If you use a 0.05 level of significance in a two-tail hypothesis test, what decision will you make if $Z_{STAT} = -0.76$?

9.2 If you use a 0.05 level of significance in a two-tail hypothesis test, what decision will you make if $Z_{STAT} = +2.21$?

9.3 If you use a 0.10 level of significance in a two-tail hypothesis test, what is your decision rule for rejecting a null hypothesis that the population mean equals 500 if you use the Z test?

9.4 If you use a 0.01 level of significance in a two-tail hypothesis test, what is your decision rule for rejecting $H_0: \mu = 12.5$ if you use the Z test?

9.5 What is your decision in Problem 9.4 if $Z_{STAT} = -2.61$?

9.6 What is the *p*-value if, in a two-tail hypothesis test, $Z_{STAT} = +2.00$?

9.7 In Problem 9.6, what is your statistical decision if you test the null hypothesis at the 0.10 level of significance?

9.8 What is the *p*-value if, in a two-tail hypothesis test, $Z_{STAT} = -1.38$?

APPLYING THE CONCEPTS

9.9 In the U.S. legal system, a defendant is presumed innocent until proven guilty. Consider a null hypothesis, H_0 , that a defendant is innocent, and an alternative hypothesis, H_1 , that the defendant is guilty. A jury has two possible decisions: Convict the defendant (i.e., reject the null hypothesis) or do not convict the defendant (i.e., do not reject the null hypothesis). Explain the meaning of the risks of committing either a Type I or Type II error in this example.

9.10 Suppose the defendant in Problem 9.9 is presumed guilty until proven innocent. How do the null and alternative hypotheses differ from those in Problem 9.9? What are the meanings of the risks of committing either a Type I or Type II error here?

9.11 Many consumer groups feel that the U.S. Food and Drug Administration (FDA) drug approval process is too easy and, as a result, too many drugs are approved that are later found to be unsafe. On the other hand, a number of industry lobbyists have pushed for a more lenient approval process so that pharmaceutical companies can get new drugs approved more easily and quickly. Consider a null hypothesis that a new, unapproved drug is unsafe and an alternative hypothesis that a new, unapproved drug is safe.

- Explain the risks of committing a Type I or Type II error.
- Which type of error are the consumer groups trying to avoid? Explain.
- Which type of error are the industry lobbyists trying to avoid? Explain.
- How would it be possible to lower the chances of both Type I and Type II errors?

9.12 As a result of complaints from both students and faculty about lateness, the registrar at a large university is ready to undertake a study to determine whether the scheduled break between classes should be changed. Until now, the registrar has believed that there should be 20 minutes between scheduled classes. State the null hypothesis, H_0 , and the alternative hypothesis, H_1 .

9.13 Do marketing majors at your school study more than, less than, or about the same as marketing majors at other schools? *The Washington Post* reported the results of the National Survey of

Student Engagement that found marketing majors studied an average of 12.1 hours per week. (Data extracted from “Is College Too Easy? As Study Time Falls, Debate Rises,” *The Washington Post*, May 21, 2012.) Set up a hypothesis test to try to prove that the mean number of hours studied by marketing majors at your school is different from the 12.1-hour-per-week benchmark reported by *The Washington Post*.

- State the null and alternative hypothesis.
- What is a Type I error for your test?
- What is a Type II error for your test?



9.14 The quality-control manager at a compact fluorescent light bulb (CFL) factory needs to determine whether the mean life of a large shipment of CFLs is equal to 7,500 hours. The population standard deviation is 1,000 hours. A random sample of 64 CFLs indicates a sample mean life of 7,250 hours.

- At the 0.05 level of significance, is there evidence that the mean life is different from 7,500 hours?
- Compute the *p*-value and interpret its meaning.
- Construct a 95% confidence interval estimate of the population mean life of the CFLs.
- Compare the results of (a) and (c). What conclusions do you reach?

9.15 Suppose that in Problem 9.14, the standard deviation is 1,200 hours.

- Repeat (a) through (d) of Problem 9.14, assuming a standard deviation of 1,200 hours.
- Compare the results of (a) to those of Problem 9.14.

9.16 A bottled water distributor wants to determine whether the mean amount of water contained in 1-gallon bottles purchased from a nationally known water bottling company is actually 1 gallon. You know from the water bottling company specifications that the standard deviation of the amount of water per bottle is 0.02 gallon. You select a random sample of 50 bottles, and the mean amount of water per 1-gallon bottle is 0.995 gallon.

- Is there evidence that the mean amount is different from 1.0 gallon? (Use $\alpha = 0.01$.)
- Compute the *p*-value and interpret its meaning.
- Construct a 99% confidence interval estimate of the population mean amount of water per bottle.
- Compare the results of (a) and (c). What conclusions do you reach?

9.17 Suppose that in Problem 9.16, the standard deviation is 0.012 gallon.

- Repeat (a) through (d) of Problem 9.16, assuming a standard deviation of 0.012 gallon.
- Compare the results of (a) to those of Problem 9.16.

9.2 *t* Test of Hypothesis for the Mean (σ Unknown)

In virtually all hypothesis-testing situations concerning the population mean, μ , you do not know the population standard deviation, σ . Instead, you use the sample standard deviation, S . If you assume that the population is normally distributed, the sampling distribution of the mean follows a *t* distribution with $n - 1$ degrees of freedom, and you use the ***t* test for the mean**. If the population is not normally distributed, you can still use the *t* test if the population is not too skewed and the sample size is not too small. Equation (9.2) defines the test statistic for determining the difference between the sample mean, \bar{X} , and the population mean, μ , when using the sample standard deviation, S .

t TEST FOR THE MEAN (σ UNKNOWN)

$$t_{STAT} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \quad (9.2)$$

where the t_{STAT} test statistic follows a t distribution having $n - 1$ degrees of freedom.

To illustrate the use of the t test for the mean, return to the Chapter 8 Ricknel Home Centers scenario on page 272. The business objective is to determine whether the mean amount per sales invoice is unchanged from the \$120 of the past five years. As an accountant for the company, you need to determine whether this amount has changed. In other words, the hypothesis test is used to try to determine whether the mean amount per sales invoice is increasing or decreasing.

The Critical Value Approach

 **Student Tip**
Remember, the null hypothesis uses an equal sign and the alternative hypothesis *never* uses an equal sign.

To perform this two-tail hypothesis test, you use the six-step method listed in Exhibit 9.1 on page 315.

Step 1 You define the following hypotheses:

$$\begin{aligned} H_0 &: \mu = 120 \\ H_1 &: \mu \neq 120 \end{aligned}$$

The alternative hypothesis contains the statement you are trying to prove. If the null hypothesis is rejected, then there is statistical evidence that the population mean amount per sales invoice is no longer \$120. If the statistical conclusion is “do not reject H_0 ,” then you will conclude that there is insufficient evidence to prove that the mean amount differs from the long-term mean of \$120.

Step 2 You collect the data from a sample of $n = 12$ sales invoices. You decide to use $\alpha = 0.05$.

Step 3 Because σ is unknown, you use the t distribution and the t_{STAT} test statistic. You must assume that the population of sales invoices is approximately normally distributed in order to use the t distribution because the sample size is only 12. This assumption is discussed on page 324.

Step 4 For a given sample size, n , the test statistic t_{STAT} follows a t distribution with $n - 1$ degrees of freedom. The critical values of the t distribution with $12 - 1 = 11$ degrees of freedom are found in Table E.3, as illustrated in Table 9.2 and Figure 9.6. The alternative hypothesis, $H_1 : \mu \neq 120$, has two tails. The area in the rejection region of the t distribution’s left (lower) tail is 0.025, and the area in the rejection region of the t distribution’s right (upper) tail is also 0.025.

From the t table as given in Table E.3, a portion of which is shown in Table 9.2, the critical values are ± 2.2010 . The decision rule is

 **Student Tip**
Since this is a two-tail test, the level of significance, $\alpha = 0.05$, is divided into two equal 0.025 parts, in each of the two tails of the distribution.

Reject H_0 if $t_{STAT} < -2.2010$
or if $t_{STAT} > +2.2010$;
otherwise, do not reject H_0 .

TABLE 9.2

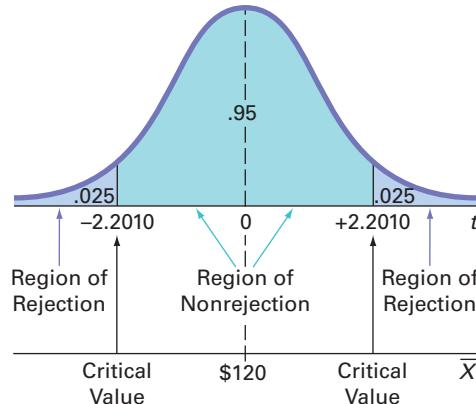
Determining the Critical Value from the *t* Table for an Area of 0.025 in Each Tail, with 11 Degrees of Freedom

Degrees of Freedom	Cumulative Probabilities					
	.75	.90	.95	.975	.99	.995
	Upper-Tail Areas					
Degrees of Freedom	.25	.10	.05	.025	.01	.005
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058

Source: Extracted from Table E.3.

FIGURE 9.6

Testing a hypothesis about the mean (σ unknown) at the 0.05 level of significance with 11 degrees of freedom



Step 5 You organize and store the data from a random sample of 12 sales invoices in **Invoices**:

108.98	152.22	111.45	110.59	127.46	107.26
93.32	91.97	111.56	75.71	128.58	135.11

Using Equations (3.1) and (3.7) on pages 102 and 109,

$$\bar{X} = \$112.85 \text{ and } S = \$20.80$$

From Equation (9.2) on page 322,

$$t_{STAT} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{112.85 - 120}{\frac{20.80}{\sqrt{12}}} = -1.1908$$

Step 6 Because $-2.2010 < t_{STAT} = -1.1908 < 2.2010$, you do not reject H_0 . You have insufficient evidence to conclude that the mean amount per sales invoice differs from \$120. The audit suggests that the mean amount per invoice has not changed.

Figure 9.7 shows the results for this test of hypothesis, as computed by Excel and Minitab.

FIGURE 9.7

Excel and Minitab results for the t test of sales invoices

A B	
1 t Test for the Hypothesis of the Mean	
2	
3	Data
4 Null Hypothesis	$\mu =$ 120
5 Level of Significance	0.05
6 Sample Size	12
7 Sample Mean	112.85
8 Sample Standard Deviation	20.8
9	
10 Intermediate Calculations	
11 Standard Error of the Mean	6.0044 =B8/SQRT(B6)
12 Degrees of Freedom	11 =B6 - 1
13 t Test Statistic	-1.1908 =(B7 - B4)/B11
14	
15 Two-Tail Test	
16 Lower Critical Value	-2.2010 =-T.INV.2T(B5, B12)
17 Upper Critical Value	2.2010 =T.INV.2T(B5, B12)
18 p-Value	0.2588 =T.DIST.2T(ABS(B13), B12)
19 Do not reject the null hypothesis	=IF(B18 < B5, "Reject the null hypothesis", "Do not reject the null hypothesis")

One-Sample T						
Test of $\mu = 120$ vs not = 120						
N	Mean	StDev	SE Mean	95% CI	T	P
12	112.85	20.80	6.00	(99.63, 126.07)	-1.19	0.259

The p-Value Approach

To perform this two-tail hypothesis test, you use the five-step method listed in Exhibit 9.2 on page 318.

Step 1–3 These steps are the same as in the critical value approach discussed on page 322.

Step 4 From the Figure 9.7 results, $t_{STAT} = -1.19$ and the p -value = 0.2588

Step 5 Because the p -value of 0.2588 is greater than $\alpha = 0.05$, you do not reject H_0 . The data provide insufficient evidence to conclude that the mean amount per sales invoice differs from \$120. The audit suggests that the mean amount per invoice has not changed. The p -value indicates that if the null hypothesis is true, the probability that a sample of 12 invoices could have a sample mean that differs by \$7.15 or more from the stated \$120 is 0.2588. In other words, if the mean amount per sales invoice is truly \$120, then there is a 25.88% chance of observing a sample mean below \$112.85 or above \$127.15.

In the preceding example, it is incorrect to state that there is a 25.88% chance that the null hypothesis is true. Remember that the p -value is a conditional probability, calculated by *assuming* that the null hypothesis is true. In general, it is proper to state the following:

If the null hypothesis is true, there is a (p -value) \times 100% chance of observing a test statistic at least as contradictory to the null hypothesis as the sample result.

Checking the Normality Assumption

You use the t test when the population standard deviation, σ , is not known and is estimated using the sample standard deviation, S . To use the t test, you assume that the data represent a random sample from a population that is normally distributed. In practice, as long as the sample size is not very small and the population is not very skewed, the t distribution provides a good approximation of the sampling distribution of the mean when σ is unknown.

There are several ways to evaluate the normality assumption necessary for using the t test. You can examine how closely the sample statistics match the normal distribution's theoretical properties. You can also construct a histogram, stem-and-leaf display, boxplot, or normal probability plot to visualize the distribution of the sales invoice amounts. For details on evaluating normality, see Section 6.3 on pages 233–236.

Figures 9.8 and 9.9 show the descriptive statistics, boxplot, and normal probability plot for the sales invoice data.

FIGURE 9.8

Excel and Minitab descriptive statistics and boxplots for the sales invoice data

	Invoice Amount
Mean	112.8508
Median	111.02
Mode	#N/A
Minimum	75.71
Maximum	152.22
Range	76.51
Variance	432.5565
Standard Deviation	20.7980
Coeff. of Variation	18.43%
Skewness	0.1336
Kurtosis	0.1727
Count	12
Standard Error	6.0039

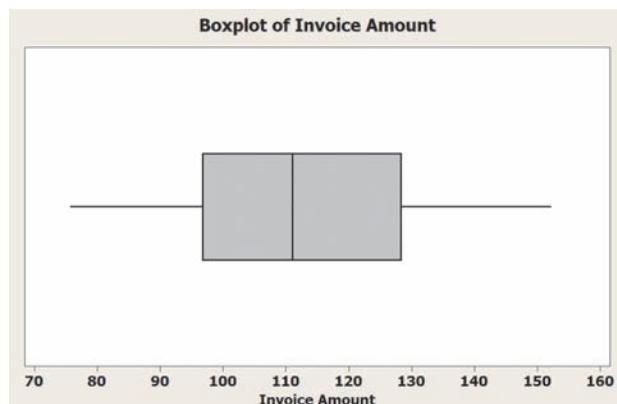
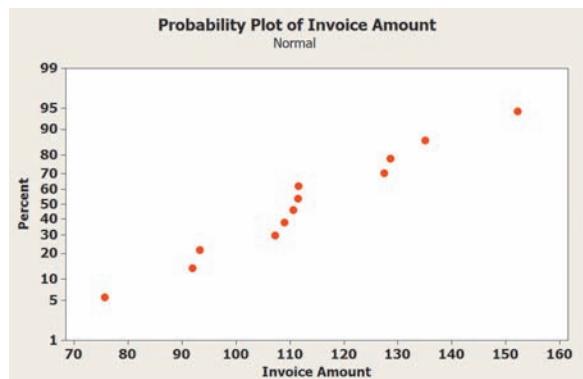
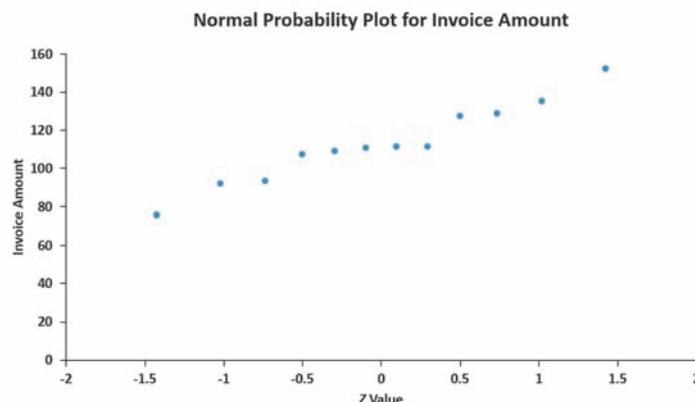


FIGURE 9.9

Excel and Minitab normal probability plots for the sales invoice data



The mean is very close to the median, and the points on the normal probability appear to be increasing approximately in a straight line. The boxplot appears to be approximately symmetrical. Thus, you can assume that the population of sales invoices is approximately normally distributed. The normality assumption is valid, and therefore the auditor's results are valid.

The *t* test is a **robust** test. A robust test does not lose power if the shape of the population departs somewhat from a normal distribution, particularly when the sample size is large enough to enable the test statistic *t* to follow the *t* distribution. However, you can reach erroneous conclusions and can lose statistical power if you use the *t* test incorrectly. If the sample size, *n*, is small (i.e., less than 30) and you cannot easily make the assumption that the underlying population is at least approximately normally distributed, then *nonparametric* testing procedures are more appropriate (see references 2 and 3).

Problems for Section 9.2

LEARNING THE BASICS

9.18 If, in a sample of $n = 16$ selected from a normal population, $\bar{X} = 56$ and $S = 12$, what is the value of t_{STAT} if you are testing the null hypothesis $H_0: \mu = 50$?

9.19 In Problem 9.18, how many degrees of freedom does the *t* test have?

9.20 In Problems 9.18 and 9.19, what are the critical values of *t* if the level of significance, α , is 0.05 and the alternative hypothesis, H_1 , is $\mu \neq 50$?

9.21 In Problems 9.18, 9.19, and 9.20, what is your statistical decision if the alternative hypothesis, H_1 , is $\mu \neq 50$?

9.22 If, in a sample of $n = 16$ selected from a left-skewed population, $\bar{X} = 65$, and $S = 21$, would you use the t test to test the null hypothesis $H_0: \mu = 60$? Discuss.

9.23 If, in a sample of $n = 160$ selected from a left-skewed population, $\bar{X} = 65$, and $S = 21$, would you use the t test to test the null hypothesis $H_0: \mu = 60$? Discuss.

APPLYING THE CONCEPTS

SELF Test **9.24** You are the manager of a restaurant for a fast-food franchise. Last month, the mean waiting time at the drive-through window for branches in your geographic region, as measured from the time a customer places an order until the time the customer receives the order, was 3.7 minutes. You select a random sample of 64 orders. The sample mean waiting time is 3.57 minutes, with a sample standard deviation of 0.8 minute.

- At the 0.05 level of significance, is there evidence that the population mean waiting time is different from 3.7 minutes?
- Because the sample size is 64, do you need to be concerned about the shape of the population distribution when conducting the t test in (a)? Explain.

9.25 A manufacturer of chocolate candies uses machines to package candies as they move along a filling line. Although the packages are labeled as 8 ounces, the company wants the packages to contain a mean of 8.17 ounces so that virtually none of the packages contain less than 8 ounces. A sample of 50 packages is selected periodically, and the packaging process is stopped if there is evidence that the mean amount packaged is different from 8.17 ounces. Suppose that in a particular sample of 50 packages, the mean amount dispensed is 8.159 ounces, with a sample standard deviation of 0.051 ounce.

- Is there evidence that the population mean amount is different from 8.17 ounces? (Use a 0.05 level of significance.)
- Determine the p -value and interpret its meaning.

9.26 A marketing researcher wants to estimate the mean savings (\$) realized by shoppers who showroom. Showrooming is the practice of inspecting products in retail stores and then purchasing the products online at a lower price. A random sample of 100 shoppers who recently purchased a consumer electronics item online after making a visit to a retail store yielded a mean savings of \$58 and a standard deviation of \$55.

- Is there evidence that the population mean savings for all showroomers who purchased a consumer electronics item is different from \$50? (Use a 0.05 level of significance.)
- Determine the p -value and interpret its meaning.

9.27 The U.S. Department of Transportation requires tire manufacturers to provide performance information on tire sidewalls to help prospective buyers make their purchasing decisions. One very important piece of information is the tread wear index, which indicates the tire's resistance to tread wear. A tire with a grade of 200 should last twice as long, on average, as a tire with a grade of 100.

A consumer organization wants to test the actual tread wear index of a brand name of tires that claims "graded 200" on the sidewall of the tire. A random sample of $n = 18$ indicates a sample mean tread wear index of 195.3 and a sample standard deviation of 21.4.

- Is there evidence that the population mean tread wear index is different from 200? (Use a 0.05 level of significance.)
- Determine the p -value and interpret its meaning.

9.28 The file **FastFood** contains the amount that a sample of fifteen customers spent for lunch (\$) at a fast-food restaurant:

7.42 6.29 5.83 6.50 8.34 9.51 7.10 6.80 5.90
4.89 6.50 5.52 7.90 8.30 9.60

- At the 0.05 level of significance, is there evidence that the mean amount spent for lunch is different from \$6.50?
- Determine the p -value in (a) and interpret its meaning.
- What assumption must you make about the population distribution in order to conduct the t test in (a) and (b)?
- Because the sample size is 15, do you need to be concerned about the shape of the population distribution when conducting the t test in (a)? Explain.

9.29 An insurance company has the business objective of reducing the amount of time it takes to approve applications for life insurance. The approval process consists of underwriting, which includes a review of the application, a medical information bureau check, possible requests for additional medical information and medical exams, and a policy compilation stage in which the policy pages are generated and sent for delivery. The ability to deliver approved policies to customers in a timely manner is critical to the profitability of this service. During a period of one month, a random sample of 27 approved policies is selected, and the total processing time, in days, is collected. These data, stored in **Insurance**, are:

73 19 16 64 28 28 31 90 60 56 31 56 22 18 45 48
17 17 17 91 92 63 50 51 69 16 17

- In the past, the mean processing time was 45 days. At the 0.05 level of significance, is there evidence that the mean processing time has changed from 45 days?
- What assumption about the population distribution is needed in order to conduct the t test in (a)?
- Construct a boxplot or a normal probability plot to evaluate the assumption made in (b).
- Do you think that the assumption needed in order to conduct the t test in (a) is valid? Explain.

9.30 The following data (in **Drink**) represent the amount of soft drink filled in a sample of 50 consecutive 2-liter bottles. The results, listed horizontally in the order of being filled, were:

2.109	2.086	2.066	2.075	2.065	2.057	2.052	2.044
2.036	2.038	2.031	2.029	2.025	2.029	2.023	2.020
2.015	2.014	2.013	2.014	2.012	2.012	2.012	2.010
2.005	2.003	1.999	1.996	1.997	1.992	1.994	1.986
1.984	1.981	1.973	1.975	1.971	1.969	1.966	1.967
1.963	1.957	1.951	1.951	1.947	1.941	1.941	1.938
1.908	1.894						

- At the 0.05 level of significance, is there evidence that the mean amount of soft drink filled is different from 2.0 liters?
- Determine the p -value in (a) and interpret its meaning.

- c. In (a), you assumed that the distribution of the amount of soft drink filled was normally distributed. Evaluate this assumption by constructing a boxplot or a normal probability plot.
- d. Do you think that the assumption needed in order to conduct the *t* test in (a) is valid? Explain.
- e. Examine the values of the 50 bottles in their sequential order, as given in the problem. Does there appear to be a pattern to the results? If so, what impact might this pattern have on the validity of the results in (a)?

9.31 One of the major measures of the quality of service provided by any organization is the speed with which it responds to customer complaints. A large family-held department store selling furniture and flooring, including carpet, had undergone a major expansion in the past several years. In particular, the flooring department had expanded from 2 installation crews to an installation supervisor, a measurer, and 15 installation crews. The store had the business objective of improving its response to complaints. The variable of interest was defined as the number of days between when the complaint was made and when it was resolved. Data were collected from 50 complaints that were made in the past year. These data, stored in **Furniture**, are:

54	5	35	137	31	27	152	2	123	81	74	27
11	19	126	110	110	29	61	35	94	31	26	5
12	4	165	32	29	28	29	26	25	1	14	13
13	10	5	27	4	52	30	22	36	26	20	23
33	68										

- a. The installation supervisor claims that the mean number of days between the receipt of a complaint and the resolution of the complaint is 20 days. At the 0.05 level of significance, is there evidence that the claim is not true (i.e., the mean number of days is different from 20)?
- b. What assumption about the population distribution is needed in order to conduct the *t* test in (a)?
- c. Construct a boxplot or a normal probability plot to evaluate the assumption made in (b).
- d. Do you think that the assumption needed in order to conduct the *t* test in (a) is valid? Explain.

9.32 A manufacturing company produces steel housings for electrical equipment. The main component part of the housing is a steel trough that is made out of a 14-gauge steel coil. It is produced using a 250-ton progressive punch press with a wipe-down operation that puts two 90-degree forms in the flat steel to make the trough. The distance from one side of the form to the other is critical because of weatherproofing in outdoor applications. The company requires that the width of the trough be between 8.31 inches and 8.61 inches. The file **Trough** contains the widths of the troughs, in inches, for a sample of $n = 49$:

8.312	8.343	8.317	8.383	8.348	8.410	8.351	8.373	8.481	8.422
8.476	8.382	8.484	8.403	8.414	8.419	8.385	8.465	8.498	8.447
8.436	8.413	8.489	8.414	8.481	8.415	8.479	8.429	8.458	8.462
8.460	8.444	8.429	8.460	8.412	8.420	8.410	8.405	8.323	8.420
8.396	8.447	8.405	8.439	8.411	8.427	8.420	8.498	8.409	

- a. At the 0.05 level of significance, is there evidence that the mean width of the troughs is different from 8.46 inches?
- b. What assumption about the population distribution is needed in order to conduct the *t* test in (a)?
- c. Evaluate the assumption made in (b).
- d. Do you think that the assumption needed in order to conduct the *t* test in (a) is valid? Explain.

9.33 One operation of a steel mill is to cut pieces of steel into parts that are used in the frame for front seats in an automobile. The steel is cut with a diamond saw and requires the resulting parts must be cut to be within ± 0.005 inch of the length specified by the automobile company. The file **Steel** contains a sample of 100 steel parts. The measurement reported is the difference, in inches, between the actual length of the steel part, as measured by a laser measurement device, and the specified length of the steel part. For example, a value of -0.002 represents a steel part that is 0.002 inch shorter than the specified length.

- a. At the 0.05 level of significance, is there evidence that the mean difference is different from 0.0 inches?
- b. Construct a 95% confidence interval estimate of the population mean. Interpret this interval.
- c. Compare the conclusions reached in (a) and (b).
- d. Because $n = 100$, do you have to be concerned about the normality assumption needed for the *t* test and *t* interval?

9.34 In Problem 3.69 on page 142, you were introduced to a tea-bag-filling operation. An important quality characteristic of interest for this process is the weight of the tea in the individual bags. The file **Teabags** contains an ordered array of the weight, in grams, of a sample of 50 tea bags produced during an 8-hour shift.

- a. Is there evidence that the mean amount of tea per bag is different from 5.5 grams? (Use $\alpha = 0.01$.)
- b. Construct a 99% confidence interval estimate of the population mean amount of tea per bag. Interpret this interval.
- c. Compare the conclusions reached in (a) and (b).

9.35 An article appearing in *The Exponent*, an independent college newspaper published by the Purdue Student Publishing Foundation, reported that the average American college student spends 1 hour (60 minutes) on Facebook daily. (Data extracted from bit.ly/QqQHow.) In order to test the validity of this statement, you select a sample of 30 Facebook users at your college. The results for the time spent on Facebook per day (in minutes) are stored in **FacebookTime**.

- a. Is there evidence that the population mean time Facebook time is different from 60 minutes? Use the *p*-value approach and a level of significance of 0.05.
- b. What assumption about the population distribution is needed in order to conduct the *t* test in (a)?
- c. Make a list of the various ways you could evaluate the assumption noted in (b).
- d. Evaluate the assumption noted in (b) and determine whether the test in (a) is valid.

9.3 One-Tail Tests

The examples of hypothesis testing in Sections 9.1 and 9.2 are called two-tail tests because the rejection region is divided into the two tails of the sampling distribution of the mean. In contrast, some hypothesis tests are one-tail tests because they require an alternative hypothesis that focuses on a *particular direction*.

One example of a one-tail hypothesis test would test whether the population mean is *less than* a specified value. One such situation involves the business problem concerning the service time at the drive-through window of a fast-food restaurant. According to *QSR* magazine, the speed with which customers are served is of critical importance to the success of the service (see bit.ly/WoJpTT). In one past study, an audit of McDonald's drive-throughs had a mean service time of 188.83 seconds, which was slower than the drive-throughs of several other fast-food chains. Suppose that McDonald's began a quality improvement effort to reduce the service time by deploying an improved drive-through service process in a sample of 25 stores. Because McDonald's would want to institute the new process in all of its stores only if the test sample saw a *decreased* drive-through time, the entire rejection region is located in the lower tail of the distribution.

The Critical Value Approach

You wish to determine whether the new drive-through process has a mean that is less than 188.83 seconds. To perform this one-tail hypothesis test, you use the six-step method listed in Exhibit 9.1 on page 315:

Step 1 You define the null and alternative hypotheses:

$$H_0: \mu \geq 188.83$$

$$H_1: \mu < 188.83$$



Student Tip

The rejection region matches the direction of the alternative hypothesis. If the alternative hypothesis contains a $<$ sign, the rejection region is in the lower tail. If the alternative hypothesis contains a $>$ sign, the rejection region is in the upper tail.

The alternative hypothesis contains the statement for which you are trying to find evidence. If the conclusion of the test is “reject H_0 ,” there is statistical evidence that the mean drive-through time is less than the drive-through time in the old process. This would be reason to change the drive-through process for the entire population of stores. If the conclusion of the test is “do not reject H_0 ,” then there is insufficient evidence that the mean drive-through time in the new process is significantly less than the drive-through time in the old process. If this occurs, there would be insufficient reason to institute the new drive-through process in the population of stores.

Step 2 You collect the data by selecting a sample of $n = 25$ stores. You decide to use $\alpha = 0.05$.

Step 3 Because σ is unknown, you use the t distribution and the t_{STAT} test statistic. You need to assume that the drive-through time is normally distributed because a sample of only 25 drive-through times is selected.

Step 4 The rejection region is entirely contained in the lower tail of the sampling distribution of the mean because you want to reject H_0 only when the sample mean is significantly less than 188.83 seconds. When the entire rejection region is contained in one tail of the sampling distribution of the test statistic, the test is called a **one-tail test**, or **directional test**. If the alternative hypothesis includes the *less than* sign, the critical value of t is negative. As shown in Table 9.3 and Figure 9.10, because the entire rejection region is in the lower tail of the t distribution and contains an area of 0.05, due to the symmetry of the t distribution, the critical value of the t test statistic with $25 - 1 = 24$ degrees of freedom is -1.7109 .

The decision rule is

Reject H_0 if $t_{STAT} < -1.7109$;

otherwise, do not reject H_0 .

TABLE 9.3

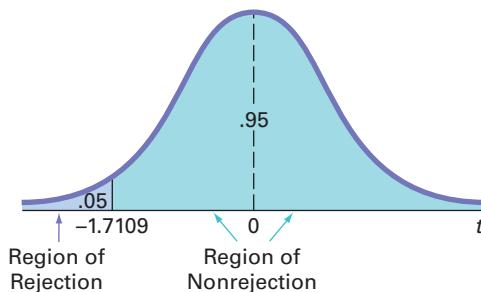
Determining the Critical Value from the *t* Table for an Area of 0.05 in the Lower Tail, with 24 Degrees of Freedom

Degrees of Freedom	Cumulative Probabilities					
	.75	.90	.95	.975	.99	.995
	Upper-Tail Areas					
.25	.10	.05	.025	.01	.005	
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
⋮	⋮	⋮	⋮	⋮	⋮	⋮
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874

Source: Extracted from Table E.3.

FIGURE 9.10

One-tail test of hypothesis for a mean (σ unknown) at the 0.05 level of significance



Step 5 From the sample of 25 stores you selected, you find that the sample mean service time at the drive-through equals 170.8 seconds and the sample standard deviation equals 21.3 seconds. Using $n = 25$, $\bar{X} = 170.8$, $S = 21.3$, and Equation (9.2) on page 322,

$$t_{STAT} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{170.8 - 188.83}{\frac{21.3}{\sqrt{25}}} = -4.2324$$

Step 6 Because $t_{STAT} = -4.2324 < -1.7109$, you reject the null hypothesis (see Figure 9.10). You conclude that the mean service time at the drive-through is less than 188.83 seconds. There is sufficient evidence to change the drive-through process for the entire population of stores.

The *p*-Value Approach

Use the five steps listed in Exhibit 9.2 on page 318 to illustrate the *t* test for the drive-through time study using the *p*-value approach:

Step 1–3 These steps are the same as was used in the critical value approach on page 328.

Step 4 $t_{STAT} = -4.2324$ (see step 5 of the critical value approach). Because the alternative hypothesis indicates a rejection region entirely in the lower tail of the sampling distribution, to compute the *p*-value, you need to find the probability that the t_{STAT} test statistic will be less than -4.2324 . Figure 9.11 on page 330 shows that the *p*-value is 0.0001 (displayed as 0.000 in Minitab).

FIGURE 9.11

Excel and Minitab t test results for the drive-through time study

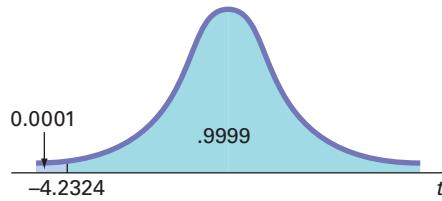
A	B
1 <i>t Test for the Hypothesis of the Mean</i>	
2	
3 <i>Data</i>	
4 Null Hypothesis $\mu =$	188.83
5 Level of Significance	0.05
6 Sample Size	25
7 Sample Mean	170.8
8 Sample Standard Deviation	21.3
9	
10 <i>Intermediate Calculations</i>	
11 Standard Error of the Mean	=B8/SQRT(B6)
12 Degrees of Freedom	=B6 - 1
13 <i>t</i> Test Statistic	=B7 - B4)/B11
14	
15 <i>Lower-Tail Test</i>	
16 Lower Critical Value	=T.INV.2T(2 * B5, B12)
17 <i>p</i> -Value	=IF(B13 < 0, E11, E12)
18 Reject the null hypothesis	=IF(B17 < B5, "Reject the null hypothesis", "Do not reject the null hypothesis")
D	
10 One-Tail Calculations	
11 T.DIST.RT value	=T.DIST.RT(ABS(B13), B12)
12 1-T.DIST.RT value	=1 - E11

C	D	E				
One-Sample T						
Test of $\mu = 188.83$ vs not = 188.83						
N	Mean	StDev	SE Mean	95% CI	T	P
25	170.80	21.30	4.26	(162.01, 179.59)	-4.23	0.000

Step 5 The p -value of 0.0001 is less than $\alpha = 0.05$ (see Figure 9.12). You reject H_0 and conclude that the mean service time at the drive-through is less than 188.83 seconds. There is sufficient evidence to change the drive-through process for the entire population of stores.

FIGURE 9.12

Determining the p -value for a one-tail test



Example 9.5 illustrates a one-tail test in which the rejection region is in the upper tail.

EXAMPLE 9.5

A One-Tail Test for the Mean

A company that manufactures chocolate bars is particularly concerned that the mean weight of a chocolate bar is not greater than 6.03 ounces. A sample of 50 chocolate bars is selected; the sample mean is 6.034 ounces, and the sample standard deviation is 0.02 ounce. Using the $\alpha = 0.01$ level of significance, is there evidence that the population mean weight of the chocolate bars is greater than 6.03 ounces?

SOLUTION Using the critical value approach, listed in Exhibit 9.1 on page 315,

Step 1 First, you define the null and alternative hypotheses:

$$H_0: \mu \leq 6.03$$

$$H_1: \mu > 6.03$$

Step 2 You collect the data from a sample of $n = 50$. You decide to use $\alpha = 0.01$.

Step 3 Because σ is unknown, you use the t distribution and the t_{STAT} test statistic.

Step 4 The rejection region is entirely contained in the upper tail of the sampling distribution of the mean because you want to reject H_0 only when the sample mean is significantly greater than 6.03 ounces. Because the entire rejection region is in the upper tail of the t distribution and contains an area of 0.01, the critical value of the t distribution with $50 - 1 = 49$ degrees of freedom is 2.4049 (see Table E.3).

The decision rule is

Reject H_0 if $t_{STAT} > 2.4049$;
otherwise, do not reject H_0 .

Step 5 From your sample of 50 chocolate bars, you find that the sample mean weight is 6.034 ounces, and the sample standard deviation is 0.02 ounces. Using $n = 50$, $\bar{X} = 6.034$, $S = 0.02$, and Equation (9.2) on page 322,

$$t_{STAT} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{6.034 - 6.03}{\frac{0.02}{\sqrt{50}}} = 1.414$$

Step 6 Because $t_{STAT} = 1.414 < 2.4049$ or the p -value (from Excel) is $0.0818 > 0.01$, you do not reject the null hypothesis. There is insufficient evidence to conclude that the population mean weight is greater than 6.03 ounces.

To perform one-tail tests of hypotheses, you must properly formulate H_0 and H_1 . A summary of the null and alternative hypotheses for one-tail tests is as follows:

- The null hypothesis, H_0 , represents the status quo or the current belief in a situation.
- The alternative hypothesis, H_1 , is the opposite of the null hypothesis and represents a research claim or specific inference you would like to prove.
- If you reject the null hypothesis, you have statistical proof that the alternative hypothesis is correct.
- If you do not reject the null hypothesis, you have failed to prove the alternative hypothesis. The failure to prove the alternative hypothesis, however, does not mean that you have proven the null hypothesis.
- The null hypothesis always refers to a specified value of the *population parameter* (such as μ), not to a *sample statistic* (such as \bar{X}).
- The statement of the null hypothesis *always* contains an equal sign regarding the specified value of the parameter (e.g., $H_0: \mu \geq 188.83$).
- The statement of the alternative hypothesis *never* contains an equal sign regarding the specified value of the parameter (e.g., $H_1: \mu < 188.83$).

Problems for Section 9.3

LEARNING THE BASICS

9.36 In a one-tail hypothesis test where you reject H_0 only in the *upper* tail, what is the p -value if $Z_{STAT} = +2.00$?

9.37 In Problem 9.36, what is your statistical decision if you test the null hypothesis at the 0.05 level of significance?

9.38 In a one-tail hypothesis test where you reject H_0 only in the *lower* tail, what is the p -value if $Z_{STAT} = -1.38$?

9.39 In Problem 9.38, what is your statistical decision if you test the null hypothesis at the 0.01 level of significance?

9.40 In a one-tail hypothesis test where you reject H_0 only in the *lower* tail, what is the p -value if $Z_{STAT} = +1.38$?

9.41 In Problem 9.40, what is the statistical decision if you test the null hypothesis at the 0.01 level of significance?

9.42 In a one-tail hypothesis test where you reject H_0 only in the *upper* tail, what is the critical value of the t -test statistic with 10 degrees of freedom at the 0.01 level of significance?

9.43 In Problem 9.42, what is your statistical decision if $t_{STAT} = +2.39$?

9.44 In a one-tail hypothesis test where you reject H_0 only in the *lower* tail, what is the critical value of the t_{STAT} test statistic with 20 degrees of freedom at the 0.01 level of significance?

9.45 In Problem 9.44, what is your statistical decision if $t_{STAT} = -1.15$?

APPLYING THE CONCEPTS

9.46 The Los Angeles County Metropolitan Transportation Authority has set a bus mechanical reliability goal of 3,900 bus miles. Bus mechanical reliability is measured specifically as the number of bus miles between mechanical road calls. Suppose a sample of 100 buses resulted in a sample mean of 3,975 bus miles and a sample standard deviation of 275 bus miles.

- Is there evidence that the population mean bus miles is more than 3,900 bus miles? (Use a 0.05 level of significance.)
- Determine the p -value and interpret its meaning.

9.47 *CarMD* reports that the cost of repairing a hybrid vehicle is falling even while typical repairs on conventional vehicles are getting more expensive. The most common hybrid repair, replacing the hybrid inverter assembly, had a mean repair cost of \$3,927 in 2012. Industry experts suspect that the cost will continue to decrease given the increase in the number of technicians who have gained expertise on fixing gas–electric engines in recent months. Suppose a sample of 100 hybrid inverter assembly repairs completed in the last month was selected. The sample mean repair cost was \$3,800 with the sample standard deviation of \$500.

- Is there evidence that the population mean cost is less than \$3,927? (Use a 0.05 level of significance.)
- Determine the *p*-value and interpret its meaning.

 **9.48** A quality improvement project was conducted with the objective of improving the wait time in a county health department (CHD) Adult Primary Care Unit (APCU). The evaluation plan included *waiting room time* as one key waiting time process measure. Waiting room time was defined as the time elapsed between requesting that the patient be seated in the waiting room and the time he or she was called to be placed in an exam room. Suppose that, initially, a targeted wait time goal of 25 minutes was set. After implementing an improvement framework and process, the quality improvement team collected data on a sample of 355 patients. In this sample, the mean wait time was 23.05 minutes, with a standard deviation of 16.83 minutes. (Data extracted from M. Michael, S. D. Schaffer, P. L. Egan, B. B. Little, and P. S. Pritchard, “Improving Wait Times and Patient Satisfaction in Primary Care,” *Journal for Healthcare Quality*, 2013, 35(2), pp. 50–60.)

- If you test the null hypothesis at the 0.01 level of significance, is there evidence that the population mean wait time is less than 25 minutes?
- Interpret the meaning of the *p*-value in this problem.

9.49 You are the manager of a restaurant that delivers pizza to college dormitory rooms. You have just changed your delivery process in an effort to reduce the mean time between the order and

completion of delivery from the current 25 minutes. A sample of 36 orders using the new delivery process yields a sample mean of 22.4 minutes and a sample standard deviation of 6 minutes.

- Using the six-step critical value approach, at the 0.05 level of significance, is there evidence that the population mean delivery time has been reduced below the previous population mean value of 25 minutes?
- At the 0.05 level of significance, use the five-step *p*-value approach.
- Interpret the meaning of the *p*-value in (b).
- Compare your conclusions in (a) and (b).

9.50 A survey of nonprofit organizations showed that online fundraising has increased in the past year. Based on a random sample of 55 nonprofit organizations, the mean one-time gift donation in the past year was \$75, with a standard deviation of \$9.

- If you test the null hypothesis at the 0.01 level of significance, is there evidence that the mean one-time gift donation is greater than \$70?
- Interpret the meaning of the *p*-value in this problem.

9.51 The population mean waiting time to check out of a supermarket has been 4 minutes. Recently, in an effort to reduce the waiting time, the supermarket has experimented with a system in which infrared cameras use body heat and in-store software to determine how many lanes should be opened. A sample of 100 customers was selected, and their mean waiting time to check out was 3.25 minutes, with a sample standard deviation of 2.7 minutes.

- At the 0.05 level of significance, using the critical value approach to hypothesis testing, is there evidence that the population mean waiting time to check out is less than 4 minutes?
- At the 0.05 level of significance, using the *p*-value approach to hypothesis testing, is there evidence that the population mean waiting time to check out is less than 4 minutes?
- Interpret the meaning of the *p*-value in this problem.
- Compare your conclusions in (a) and (b).

9.4 Z Test of Hypothesis for the Proportion

In some situations, you want to test a hypothesis about the proportion of events of interest in the population, π , rather than test the population mean. To begin, you select a random sample and compute the **sample proportion**, $p = X/n$. You then compare the value of this statistic to the hypothesized value of the parameter, π , in order to decide whether to reject the null hypothesis.

If the number of events of interest (X) and the number of events that are not of interest ($n - X$) are each at least five, the sampling distribution of a proportion approximately follows a normal distribution, and you can use the **Z test for the proportion**. Equation (9.3) defines this hypothesis test for the difference between the sample proportion, p , and the hypothesized population proportion, π .

Z TEST FOR THE PROPORTION

$$Z_{STAT} = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \quad (9.3)$$

Student Tip

Do not confuse this use of the Greek letter pi, π , to represent the population proportion with the mathematical constant that is the ratio of the circumference to a diameter of a circle—approximately 3.14159—which is also known by the same Greek letter.

where

$$p = \text{sample proportion} = \frac{X}{n} = \frac{\text{number of events of interest in the sample}}{\text{sample size}}$$

$$\pi = \text{hypothesized proportion of events of interest in the population}$$

The Z_{STAT} test statistic approximately follows a standardized normal distribution when X and $(n - X)$ are each at least 5.

Alternatively, by multiplying the numerator and denominator by n , you can write the Z_{STAT} test statistic in terms of the number of events of interest, X , as shown in Equation (9.4).

Z TEST FOR THE PROPORTION IN TERMS OF THE NUMBER OF EVENTS OF INTEREST

$$Z_{STAT} = \frac{X - n\pi}{\sqrt{n\pi(1 - \pi)}} \quad (9.4)$$

The Critical Value Approach

To illustrate the Z test for a proportion, consider a survey that sought to determine whether adults could stop thinking about work while on vacation. (Data extracted from “Can You Stop Thinking About Work on Your Vacation?” *USA Today*, October 5, 2011, p. 1A.) Of 1,000 adults, 320 said that they were unable to stop thinking about work while on vacation. Suppose that a survey conducted in the previous year indicated that 30% of adults were unable to stop thinking about work while on vacation. Is there evidence that the percentage of adults who were unable to stop thinking about work while on vacation has changed from the previous year? To investigate this question, the null and alternative hypotheses are follows:

$H_0 : \pi = 0.30$ (i.e., the proportion of adults who were unable to stop thinking about work while on vacation has not changed from the previous year)

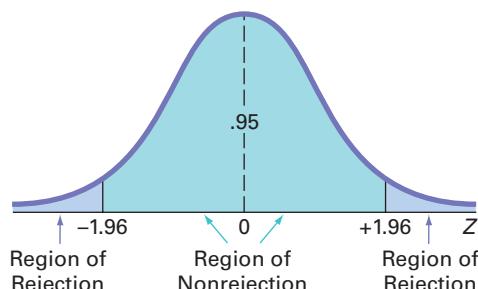
$H_1 : \pi \neq 0.30$ (i.e., the proportion of adults who were unable to stop thinking about work while on vacation has changed from the previous year)

Because you are interested in determining whether the population proportion of adults who were unable to stop thinking about work while on vacation has changed from 0.30 in the previous year, you use a two-tail test. If you select the $\alpha = 0.05$ level of significance, the rejection and nonrejection regions are set up as in Figure 9.13, and the decision rule is

Reject H_0 if $Z_{STAT} < -1.96$ or if $Z_{STAT} > +1.96$;
otherwise, do not reject H_0 .

FIGURE 9.13

Two-tail test of hypothesis for the proportion at the 0.05 level of significance



Because 320 of the 1,000 adults stated that they were unable to stop thinking about work while on vacation,

$$p = \frac{320}{1,000} = 0.32$$

Since $X = 320$ and $n - X = 680$, each > 5 , using Equation (9.3),

$$Z_{STAT} = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} = \frac{0.32 - 0.30}{\sqrt{\frac{0.30(1 - 0.30)}{1,000}}} = \frac{0.02}{0.0145} = 1.3801$$

or, using Equation (9.4),

$$Z_{STAT} = \frac{X - n\pi}{\sqrt{n\pi(1 - \pi)}} = \frac{320 - (1,000)(0.30)}{\sqrt{1,000(0.30)(0.70)}} = \frac{20}{14.4914} = 1.3801$$

Because $Z_{STAT} = 1.3801 < 1.96$, you do not reject H_0 . There is insufficient evidence that the population proportion of all adults who were unable to stop thinking about work on vacation has changed from 0.30 in the previous year. Figure 9.14 presents the Excel and Minitab results for these data.

FIGURE 9.14

Excel and Minitab results for the Z test for whether the proportion of adults who were unable to stop thinking about work while on vacation has changed from the previous year

A		B	Test and CI for One Proportion					
1	Z Test of Hypothesis for the Proportion		Test of $p = 0.3$ vs $p \neq 0.3$					
2								
3	Data		Sample	X	N	Sample p	95% CI	Z-Value
4	Null Hypothesis	$\pi =$	1	320	1000	0.320000	(0.291088, 0.348912)	1.38
5	Level of Significance							0.168
6	Number of Items of Interest	320						
7	Sample Size	1000						
8			Using the normal approximation.					
9	Intermediate Calculations							
10	Sample Proportion	0.3200	=B6/B7					
11	Standard Error	0.0145	=SQRT(B4 * (1 - B4)/B7)					
12	Z Test Statistic	1.3801	=B10 - B4)/B11					
13								
14	Two-Tail Test							
15	Lower Critical Value	-1.9600	=NORM.S.INV(B5/2)					
16	Upper Critical value	1.9600	=NORM.S.INV(1 - B5/2)					
17	p-Value	0.1675	=2 * (1 - NORM.S.DIST(ABS(B12), TRUE))					
18	Do not reject the null hypothesis		=IF(B17 < B5, "Reject the null hypothesis", "Do not reject the null hypothesis")					

The p-Value Approach

As an alternative to the critical value approach, you can compute the p -value. For this two-tail test in which the rejection region is located in the lower tail and the upper tail, you need to find the area below a Z value of -1.3801 and above a Z value of $+1.3801$. Figure 9.14 reports a p -value of 0.1675. Because this value is greater than the selected level of significance ($\alpha = 0.05$), you do not reject the null hypothesis.

Example 9.6 illustrates a one-tail test for a proportion.

EXAMPLE 9.6

Testing a Hypothesis for a Proportion

In addition to the business problem of the speed of service at the drive-through, fast-food chains want to fill orders correctly. The same audit that reported that McDonald's had a drive-through service time of 188.83 seconds also reported that McDonald's filled 90.9% of its drive-through orders correctly (see www.qsrmagazine.com/content/2012-qsr-drive-thru-study-order=accuracy). Suppose that McDonald's begins a quality improvement effort to ensure that orders at the drive-through are filled correctly. The business problem is defined as determining whether the new process can increase the percentage of orders filled correctly.

Data are collected from a sample of 400 orders using the new process. The results indicate that 378 orders were filled correctly. At the 0.01 level of significance, can you conclude that the new process has increased the proportion of orders filled correctly?

SOLUTION The null and alternative hypotheses are

$H_0: \pi \leq 0.909$ (i.e., the population proportion of orders filled correctly using the new process is less than or equal to 0.909)

$H_1: \pi > 0.909$ (i.e., the population proportion of orders filled correctly using the new process is greater than 0.909)

Since $X = 374$ and $n - X = 26$, both > 5 , using Equation (9.3) on page 332,

$$p = \frac{X}{n} = \frac{378}{400} = 0.945$$

$$Z_{STAT} = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} = \frac{0.945 - 0.909}{\sqrt{\frac{0.909(1 - 0.909)}{400}}} = \frac{0.036}{0.0144} = 2.5034$$

The p -value (computed by Excel) for $Z_{STAT} > 2.5034$ is 0.0062.

Using the critical value approach, you reject H_0 if $Z_{STAT} > 2.33$. Using the p -value approach, you reject H_0 if the p -value < 0.01 . Because $Z_{STAT} = 2.5034 > 2.33$ or the p -value $= 0.0062 < 0.01$, you reject H_0 . You have evidence that the new process has increased the proportion of correct orders above 0.909 or 90.9%.

Problems for Section 9.4

LEARNING THE BASICS

9.52 If, in a random sample of 400 items, 88 are defective, what is the sample proportion of defective items?

9.53 In Problem 9.52, if the null hypothesis is that 20% of the items in the population are defective, what is the value of Z_{STAT} ?

9.54 In Problems 9.52 and 9.53, suppose you are testing the null hypothesis $H_0: \pi = 0.20$ against the two-tail alternative hypothesis $H_1: \pi \neq 0.20$ and you choose the level of significance $\alpha = 0.05$. What is your statistical decision?

APPLYING THE CONCEPTS

9.55 The U.S. Department of Education reports that 40% of full-time college students are employed while attending college. (Data extracted from National Center for Education Statistics, *The Condition of Education 2012*, nces.ed.gov/pubs2012/2012045.pdf.) A recent survey of 60 full-time students at a university found that 25 were employed.

- a. Use the five-step p -value approach to hypothesis testing and a 0.05 level of significance to determine whether the proportion of full-time students at the university is different from the national norm of 0.40.
- b. Assume that the study found that 32 of the 60 full-time students were employed and repeat (a). Are the conclusions the same?

9.56 The worldwide market share for the Mozilla Firefox web browser was 20.3% in a recent month. (Data extracted from netmarketshare.com.) Suppose that you decide to select a sample

of 100 students at your university and you find that 25 use the Mozilla Firefox web browser.

- a. Use the five-step p -value approach to try to determine whether there is evidence that the market share for the Mozilla Firefox web browser at your university is greater than the worldwide market share of 20.3%. (Use the 0.05 level of significance.)
- b. Suppose that the sample size is $n = 400$, and you find that 25% of the sample of students at your university (100 out of 400) use the Mozilla Firefox web browser. Use the five-step p -value approach to try to determine whether there is evidence that the market share for the Mozilla Firefox web browser at your university is greater than the worldwide market share of 20.3%. (Use the 0.05 level of significance.)
- c. Discuss the effect that sample size has on hypothesis testing.
- d. What do you think are your chances of rejecting any null hypothesis concerning a population proportion if a sample size of $n = 20$ is used?

9.57 One of the issues facing organizations is increasing diversity throughout an organization. One of the ways to evaluate an organization's success at increasing diversity is to compare the percentage of employees in the organization in a particular position with a specific background to the percentage in a particular position with that specific background in the general workforce. Recently, a large academic medical center determined that 9 of 17 employees in a particular position were female, whereas 55% of the employees for this position in the general workforce were female. At the 0.05 level of significance, is there evidence that the

proportion of females in this position at this medical center is different from what would be expected in the general workforce?

 **9.58** Of 801 surveyed active LinkedIn members, 328 reported that they are planning to spend at least \$1,000 on consumer electronics in the coming year. (Data extracted from bit.ly/RITfU.) At the 0.05 level of significance, is there evidence that the proportion of all LinkedIn members who plan to spend at least \$1,000 on consumer electronics in the coming year is different from 35%?

9.59 A cellphone provider has the business objective of wanting to estimate the proportion of subscribers who would upgrade to a new cellphone with improved features if it were made available at a substantially reduced cost. Data are collected from a random sample of 500 subscribers. The results indicate that 135 of the subscribers would upgrade to a new cellphone at a reduced cost.

- At the 0.05 level of significance, is there evidence that more than 20% of the customers would upgrade to a new cellphone at a reduced cost?
- How would the manager in charge of promotional programs concerning residential customers use the results in (a)?

9.60 Actuation Consulting and Enterprise Agility recently conducted a global survey of product teams with the goal of better understanding the dynamics of product team performance and uncovering the practices that make these teams successful. One question posed was “In which of the following ways does your organization support aligning members of a core product team?” Global respondents were offered five choices. (Data extracted from www.actuationconsultingllc.com/blog/?p=285.) The most common response (31%) was “shared organizational goals and objectives linking the team.” Suppose another study is conducted to check the validity of this result, with the goal of proving that the percentage is less than 31%.

- State the null and research hypotheses.
- A sample of 100 organizations is selected, and results indicate that 28 organizations respond that “shared organizational goals and objectives linking the team” is the supported driver of alignment. Use either the six-step critical value hypothesis-testing approach or the five-step *p*-value approach to determine at the 0.05 level of significance whether there is evidence that the percentage is less than 31%.

9.5 Potential Hypothesis-Testing Pitfalls and Ethical Issues

To this point, you have studied the fundamental concepts of hypothesis testing. You have used hypothesis testing to analyze differences between sample statistics and hypothesized population parameters in order to make business decisions concerning the underlying population characteristics. You have also learned how to evaluate the risks involved in making these decisions.

When planning to carry out a hypothesis test based on a survey, research study, or designed experiment, you must ask several questions to ensure that you use proper methodology. You need to raise and answer questions such as the following in the planning stage:

- What is the goal of the survey, study, or experiment? How can you translate the goal into a null hypothesis and an alternative hypothesis?
- Is the hypothesis test a two-tail test or one-tail test?
- Can you select a random sample from the underlying population of interest?
- What types of data will you collect in the sample? Are the variables numerical or categorical?
- At what level of significance should you conduct the hypothesis test?
- Is the intended sample size large enough to achieve the desired power of the test for the level of significance chosen?
- What statistical test procedure should you use and why?
- What conclusions and interpretations can you reach from the results of the hypothesis test?

Failing to consider these questions early in the planning process can lead to biased or incomplete results. Proper planning can help ensure that the statistical study will provide objective information needed to make good business decisions.

Statistical Significance Versus Practical Significance

You need to make a distinction between the existence of a statistically significant result and its practical significance in a field of application. Sometimes, due to a very large sample size, you may get a result that is statistically significant but has little practical significance. For example, suppose that prior to a national marketing campaign focusing on a series of expensive television commercials, you believe that the proportion of people who recognize

your brand is 0.30. At the completion of the campaign, a survey of 20,000 people indicates that 6,168 recognized your brand. A one-tail test trying to prove that the proportion is now greater than 0.30 results in a p -value of 0.0047, and the correct statistical conclusion is that the proportion of consumers recognizing your brand name has now increased. Was the campaign successful? The result of the hypothesis test indicates a statistically significant increase in brand awareness, but is this increase practically important? The population proportion is now estimated at $6,168/20,000 = 0.3084 = 0.3084$ or 30.84%. This increase is less than 1% more than the hypothesized value of 30%. Did the large expenses associated with the marketing campaign produce a result with a meaningful increase in brand awareness? Because of the minimal real-world impact that an increase of less than 1% has on the overall marketing strategy and the huge expenses associated with the marketing campaign, you should conclude that the campaign was not successful. On the other hand, if the campaign increased brand awareness from 30% to 50%, you would be inclined to conclude that the campaign was successful.

Statistical Insignificance Versus Importance

In contrast to the issue of the practical significance of a statistically significant result is the situation in which an important result may not be statistically significant. In a recent case (see reference 1), the U.S. Supreme Court ruled that companies cannot rely solely on whether the result of a study is significant when determining what they communicate to investors. In some situations (see reference 6), the lack of a large enough sample size may result in a nonsignificant result when in fact an important difference does exist. A study that compared male and female entrepreneurship rates globally and within Massachusetts found a significant difference globally but not within Massachusetts, even though the entrepreneurship rates for females and for males in the two geographic areas were similar (8.8% for males in Massachusetts as compared to 8.4% globally; 5% for females in both geographic areas). The difference was due to the fact that the global sample size was 20 times larger than the Massachusetts sample size.

Reporting of Findings

In conducting research, you should document both good and bad results. You should not just report the results of hypothesis tests that show statistical significance but omit those for which there is insufficient evidence in the findings. In instances in which there is insufficient evidence to reject H_0 , you must make it clear that this does not prove that the null hypothesis is true. What the result indicates is that with the sample size used, there is not enough information to *disprove* the null hypothesis.

Ethical Issues

You need to distinguish between poor research methodology and unethical behavior. Ethical considerations arise when the hypothesis-testing process is manipulated. Some of the areas where ethical issues can arise include the use of human subjects in experiments, the data collection method, the type of test (one-tail or two-tail test), the choice of the level of significance, the cleansing and discarding of data, and the failure to report pertinent findings.

9.6 Power of a Test

The power of a hypothesis test is the probability that you correctly reject a false null hypothesis. The power of a test is affected by the level of significance, the sample size, and whether the test is one-tail or two-tail. The **Section 9.6 online topic** discusses the concept of the power of a test.

USING STATISTICS

Significant Testing at Oxford Cereals, Revisited

As the plant operations manager for Oxford Cereals, you were responsible for the cereal-filling process. It was your responsibility to adjust the process when the mean fill weight in the population of boxes deviated from the company specification of 368 grams. You chose to conduct a hypothesis test.

You determined that the null hypothesis should be that the population mean fill was 368 grams. If the mean weight of the sampled boxes was sufficiently above or below the expected 368-gram mean specified by Oxford Cereals, you would reject the null hypothesis in favor of the alternative hypothesis that the mean fill was different from 368 grams. If this happened, you would stop production and take whatever action was necessary to correct the problem. If the null hypothesis was not rejected, you would continue to believe in the status quo—that the process was working correctly—and therefore take no corrective action.

Before proceeding, you considered the risks involved with hypothesis tests. If you rejected a true null hypothesis,

you would make a Type I error and conclude that the population mean fill was not 368 when it actually was 368 grams.

This error would result in adjusting the filling process even though the process was working properly. If you did not reject a false null hypothesis, you would make a Type II error and conclude that the population mean fill was 368 grams when it actually was not 368 grams. Here, you would allow the process to continue without adjustment even though the process was not working properly.

After collecting a random sample of 25 cereal boxes, you used either the six-step critical value approach or the five-step *p*-value approach to hypothesis testing. Because the test statistic fell into the nonrejection region, you did not reject the null hypothesis. You concluded that there was insufficient evidence to prove that the mean fill differed from 368 grams. No corrective action on the filling process was needed.



Shutterstock

SUMMARY

This chapter presented the foundation of hypothesis testing. You learned how to perform tests on the population mean and on the population proportion. The chapter developed both the critical value approach and the *p*-value approach to hypothesis testing.

In deciding which test to use, you should ask the following question: Does the test involve a numerical variable or a categorical variable? If the test involves a numerical variable, you use the *t* test for the mean. If the test involves a categorical variable, you use the *Z* test for the proportion. Table 9.4 lists the hypothesis tests covered in the chapter.

TABLE 9.4

Summary of Topics in Chapter 9

TYPE OF ANALYSIS	TYPE OF DATA	
	Numerical	Categorical
Hypothesis test concerning a single parameter	<i>Z</i> test of hypothesis for the mean (Section 9.1) <i>t</i> test of hypothesis for the mean (Section 9.2)	<i>Z</i> test of hypothesis for the proportion (Section 9.4)

REFERENCES

1. Bialik, C. "Making a Stat Less Significant." *The Wall Street Journal*, April 2, 2011, A5.
2. Bradley, J. V. *Distribution-Free Statistical Tests*. Upper Saddle River, NJ: Prentice Hall, 1968.
3. Daniel, W. *Applied Nonparametric Statistics*, 2nd ed. Boston: Houghton Mifflin, 1990.
4. Microsoft Excel 2013. Redmond, WA: Microsoft Corp., 2012.
5. Minitab Release 16. State College, PA: Minitab, Inc., 2010.
6. Seaman, J., and E. Allen. "Not Significant, But Important?" *Quality Progress*, August 2011, 57–59.

KEY EQUATIONS

Z Test for the Mean (σ Known)

$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad (9.1)$$

t Test for the Mean (σ Unknown)

$$t_{STAT} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \quad (9.2)$$

Z Test for the Proportion

$$Z_{STAT} = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \quad (9.3)$$

Z Test for the Proportion in Terms of the Number of Events of Interest

$$Z_{STAT} = \frac{X - n\pi}{\sqrt{n\pi(1 - \pi)}} \quad (9.4)$$

KEY TERMS

alternative hypothesis (H_1) 309
 β risk 312
confidence coefficient 313
critical value 311
directional test 328
hypothesis testing 309
level of significance (α) 312
null hypothesis (H_0) 309

one-tail test 328
p-value 317
power of a statistical test 313
region of nonrejection 311
region of rejection 311
robust 325
sample proportion 332
t test for the mean 321

test statistic 311
two-tail test 314
Type I error 312
Type II error 312
Z test for the mean 314
Z test for the proportion 332

CHECKING YOUR UNDERSTANDING

9.61 What is the difference between a null hypothesis, H_0 , and an alternative hypothesis, H_1 ?

9.62 What is the difference between a Type I error and a Type II error?

9.63 What is meant by the power of a test?

9.64 What is the difference between a one-tail test and a two-tail test?

9.65 What is meant by a *p*-value?

9.66 How can a confidence interval estimate for the population mean provide conclusions for the corresponding two-tail hypothesis test for the population mean?

9.67 What is the six-step critical value approach to hypothesis testing?

9.68 What is the five-step *p*-value approach to hypothesis testing?

CHAPTER REVIEW PROBLEMS

9.69 In hypothesis testing, the common level of significance is $\alpha = 0.05$. Some might argue for a level of significance greater than 0.05. Suppose that web designers tested the proportion of potential web page visitors with a preference for a new web design over the existing web design. The null hypothesis was that the population proportion of web page visitors preferring the new design was 0.50, and the alternative hypothesis was that it was not equal to 0.50. The *p*-value for the test was 0.20.

- a. State, in statistical terms, the null and alternative hypotheses for this example.
- b. Explain the risks associated with Type I and Type II errors in this case.
- c. What would be the consequences if you rejected the null hypothesis for a *p*-value of 0.20?

- d. What might be an argument for raising the value of α ?
- e. What would you do in this situation?
- f. What is your answer in (e) if the *p*-value equals 0.12? What if it equals 0.06?

9.70 Financial institutions utilize prediction models to predict bankruptcy. One such model is the Altman Z-score model, which uses multiple corporate income and balance sheet values to measure the financial health of a company. If the model predicts a low Z-score value, the firm is in financial stress and is predicted to go bankrupt within the next two years. If the model predicts a moderate or high Z-score value, the firm is financially healthy and is predicted to be a non-bankrupt firm (see pages.stern.nyu.edu/~ealtman/Zscores.pdf). This decision-making procedure can be expressed in the hypothesis-testing framework. The null

hypothesis is that a firm is predicted to be a non-bankrupt firm. The alternative hypothesis is that the firm is predicted to be a bankrupt firm.

- Explain the risks associated with committing a Type I error in this case.
- Explain the risks associated with committing a Type II error in this case.
- Which type of error do you think executives want to avoid? Explain.
- How would changes in the model affect the probabilities of committing Type I and Type II errors?

9.71 The Pew Research Center conducted a survey of adults, aged 18 years and older, that included 1,954 cellphone owners. The survey found that 1,016 of adult cellphone owners use their phone while watching TV. (Data extracted from “The Rise of the Connected Viewer,” *Pew Internet & American Life Project Report*, July 17, 2012, bit.ly/Q27WND). The authors of the article imply that the survey proves that more than half of all adult cellphone users use their phone while watching TV.

- Use the five-step p -value approach to hypothesis testing and a 0.05 level of significance to try to prove that more than half of all adult cellphone users use their phone while watching TV.
- Based on your result in (a), is the claim implied by the authors valid?
- Suppose the survey found that 1,000 of adult cellphone owners use their phone while watching TV. Repeat parts (a) and (b).
- Compare the results of (b) and (c).

9.72 The owner of a specialty coffee shop wants to study coffee purchasing habits of customers at her shop. She selects a random sample of 60 customers during a certain week, with the following results:

- The amount spent was $\bar{X} = \$7.25$, $S = \$1.75$
 - Thirty-one customers say they “definitely will” recommend the specialty coffee shop to family and friends.
- At the 0.05 level of significance, is there evidence that the population mean amount spent was different from \$6.50?
 - Determine the p -value in (a).
 - At the 0.05 level of significance, is there evidence that more than 50% of all the customers say they “definitely will” recommend the specialty coffee shop to family and friends?
 - What is your answer to (a) if the sample mean equals \$6.25?
 - What is your answer to (c) if 39 customers say they “definitely will” recommend the specialty coffee shop to family and friends?

9.73 An auditor for a government agency was assigned the task of evaluating reimbursement for office visits to physicians paid by Medicare. The audit was conducted on a sample of 75 of the reimbursements, with the following results:

- In 12 of the office visits, there was an incorrect amount of reimbursement.
 - The amount of reimbursement was $\bar{X} = \$93.70$, $S = \$34.55$.
- At the 0.05 level of significance, is there evidence that the population mean reimbursement was less than \$100?
 - At the 0.05 level of significance, is there evidence that the proportion of incorrect reimbursements in the population was greater than 0.10?
 - Discuss the underlying assumptions of the test used in (a).
 - What is your answer to (a) if the sample mean equals \$90?
 - What is your answer to (b) if 15 office visits had incorrect reimbursements?

9.74 A bank branch located in a commercial district of a city has the business objective of improving the process for serving customers during the noon-to-1:00 p.m. lunch period. The waiting time (defined as the time the customer enters the line until he or she reaches the teller window) of a random sample of 15 customers is collected, and the results are organized and stored in **Bank1**. These data are:

4.21	5.55	3.02	5.13	4.77	2.34	3.54	3.20
4.50	6.10	0.38	5.12	6.46	6.19	3.79	

- At the 0.05 level of significance, is there evidence that the population mean waiting time is less than 5 minutes?
- What assumption about the population distribution is needed in order to conduct the t test in (a)?
- Construct a boxplot or a normal probability plot to evaluate the assumption made in (b).
- Do you think that the assumption needed in order to conduct the t test in (a) is valid? Explain.
- As a customer walks into the branch office during the lunch hour, she asks the branch manager how long she can expect to wait. The branch manager replies, “Almost certainly not longer than 5 minutes.” On the basis of the results of (a), evaluate this statement.

9.75 A manufacturing company produces electrical insulators. If the insulators break when in use, a short circuit is likely to occur. To test the strength of the insulators, destructive testing is carried out to determine how much force is required to break the insulators. Force is measured by observing the number of pounds of force applied to the insulator before it breaks. The following data (stored in **Force**) are from 30 insulators subjected to this testing:

1,870	1,728	1,656	1,610	1,634	1,784	1,522	1,696	1,592	1,662
1,866	1,764	1,734	1,662	1,734	1,774	1,550	1,756	1,762	1,866
1,820	1,744	1,788	1,688	1,810	1,752	1,680	1,810	1,652	1,736

- At the 0.05 level of significance, is there evidence that the population mean force required to break the insulator is greater than 1,500 pounds?
- What assumption about the population distribution is needed in order to conduct the t test in (a)?
- Construct a histogram, boxplot, or normal probability plot to evaluate the assumption made in (b).
- Do you think that the assumption needed in order to conduct the t test in (a) is valid? Explain.

9.76 An important quality characteristic used by the manufacturer of Boston and Vermont asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and coloring purposes to fall off the shingles, resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and, based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet are calculated. The company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet. The file **Moisture** includes 36 measurements

(in pounds per 100 square feet) for Boston shingles and 31 for Vermont shingles.

- For the Boston shingles, is there evidence at the 0.05 level of significance that the population mean moisture content is less than 0.35 pound per 100 square feet?
- Interpret the meaning of the *p*-value in (a).
- For the Vermont shingles, is there evidence at the 0.05 level of significance that the population mean moisture content is less than 0.35 pound per 100 square feet?
- Interpret the meaning of the *p*-value in (c).
- What assumption about the population distribution is needed in order to conduct the *t* tests in (a) and (c)?
- Construct histograms, boxplots, or normal probability plots to evaluate the assumption made in (a) and (c).
- Do you think that the assumption needed in order to conduct the *t* tests in (a) and (c) is valid? Explain.

9.77 Studies conducted by the manufacturer of Boston and Vermont asphalt shingles have shown product weight to be a major factor in the customer's perception of quality. Moreover, the weight represents the amount of raw materials being used and is therefore very important to the company from a cost standpoint. The last stage of the assembly line packages the shingles before the packages are placed on wooden pallets. Once a pallet is full (a pallet for most brands holds 16 squares of shingles), it is weighed, and the measurement is recorded. The file **Pallet** contains the weight (in pounds) from a sample of 368 pallets of Boston shingles and 330 pallets of Vermont shingles.

- For the Boston shingles, is there evidence at the 0.05 level of significance that the population mean weight is different from 3,150 pounds?
- Interpret the meaning of the *p*-value in (a).
- For the Vermont shingles, is there evidence at the 0.05 level of significance that the population mean weight is different from 3,700 pounds?
- Interpret the meaning of the *p*-value in (c).

- In (a) through (d), do you have to be concerned with the normality assumption? Explain.

9.78 The manufacturer of Boston and Vermont asphalt shingles provides its customers with a 20-year warranty on most of its products. To determine whether a shingle will last through the warranty period, accelerated-life testing is conducted at the manufacturing plant. Accelerated-life testing exposes the shingle to the stresses it would be subject to in a lifetime of normal use in a laboratory setting via an experiment that takes only a few minutes to conduct. In this test, a shingle is repeatedly scraped with a brush for a short period of time, and the shingle granules removed by the brushing are weighed (in grams). Shingles that experience low amounts of granule loss are expected to last longer in normal use than shingles that experience high amounts of granule loss. The file **Granule** contains a sample of 170 measurements made on the company's Boston shingles and 140 measurements made on Vermont shingles.

- For the Boston shingles, is there evidence at the 0.05 level of significance that the population mean granule loss is different from 0.30 grams?
- Interpret the meaning of the *p*-value in (a).
- For the Vermont shingles, is there evidence at the 0.05 level of significance that the population mean granule loss is different from 0.30 grams?
- Interpret the meaning of the *p*-value in (c).
- In (a) through (d), do you have to be concerned with the normality assumption? Explain.

REPORT WRITING EXERCISE

9.79 Referring to the results of Problems 9.76 through 9.78 concerning Boston and Vermont shingles, write a report that evaluates the moisture level, weight, and granule loss of the two types of shingles.

CASES FOR CHAPTER 9

Managing Ashland MultiComm Services

Continuing its monitoring of the upload speed first described in the Chapter 6 Managing Ashland MultiComm Services case on page 245, the technical operations department wants to ensure that the mean target upload speed for all Internet service subscribers is at least 0.97 on a standard scale in which the target value is 1.0. Each day, upload speed was measured 50 times, with the following results (stored in **AMS9**).

0.854 1.023 1.005 1.030 1.219 0.977 1.044 0.778 1.122 1.114
 1.091 1.086 1.141 0.931 0.723 0.934 1.060 1.047 0.800 0.889
 1.012 0.695 0.869 0.734 1.131 0.993 0.762 0.814 1.108 0.805
 1.223 1.024 0.884 0.799 0.870 0.898 0.621 0.818 1.113 1.286
 1.052 0.678 1.162 0.808 1.012 0.859 0.951 1.112 1.003 0.972

- Compute the sample statistics and determine whether there is evidence that the population mean upload speed is less than 0.97.
- Write a memo to management that summarizes your conclusions.

Digital Case

Apply your knowledge about hypothesis testing in this Digital Case, which continues the cereal-fill-packaging dispute first discussed in the Digital Case from Chapter 7.

In response to the negative statements made by the Concerned Consumers About Cereal Cheaters (CCACC) in the Chapter 7 Digital Case, Oxford Cereals recently conducted an experiment concerning cereal packaging. The company claims that the results of the experiment refute the CCACC allegations that Oxford Cereals has been cheating consumers by packaging cereals at less than labeled weights.

Open **OxfordCurrentNews.pdf**, a portfolio of current news releases from Oxford Cereals. Review the relevant

press releases and supporting documents. Then answer the following questions:

1. Are the results of the experiment valid? Why or why not? If you were conducting the experiment, is there anything you would change?
2. Do the results support the claim that Oxford Cereals is not cheating its customers?
3. Is the claim of the Oxford Cereals CEO that many cereal boxes contain *more* than 368 grams surprising? Is it true?
4. Could there ever be a circumstance in which the results of the Oxford Cereals experiment *and* the CCACC's results are both correct? Explain.

Sure Value Convenience Stores

You work in the corporate office for a nationwide convenience store franchise that operates nearly 10,000 stores. The per-store daily customer count (i.e., the mean number of customers in a store in one day) has been steady, at 900, for some time. To increase the customer count, the chain is considering cutting prices for coffee beverages. The small size will now be \$0.59 instead of \$0.99, and the medium size will be \$0.69 instead of \$1.19. Even with this reduction in price, the chain will have a 40% gross margin on coffee.

To test the new initiative, the chain has reduced coffee prices in a sample of 34 stores, where customer counts have been running almost exactly at the national average of 900. After four weeks, the sample stores stabilize at a mean customer count of 974 and a standard deviation of 96. This increase seems like a substantial amount to you, but it also seems like a pretty small sample. Is there statistical evidence that reducing coffee prices is a good strategy for increasing the mean customer count? Be prepared to explain your conclusion.

CHAPTER 9 EXCEL GUIDE

EG9.1 FUNDAMENTALS of HYPOTHESIS-TESTING METHODOLOGY

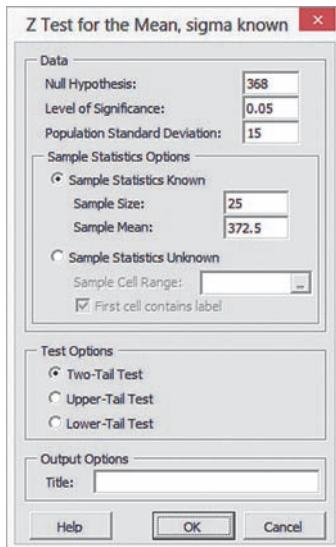
Key Technique Use the **NORM.S.INV** function to compute the lower and upper critical values and use **NORM.S.DIST (absolute value of the Z test statistic, True)** as part of a formula to compute the *p*-value. Use an **IF** function (see Appendix Section F.4) to determine whether to display a rejection or nonrejection message.

Example Perform the Figure 9.5 two-tail *Z* test for the mean for the cereal-filling example shown on page 318.

PHStat Use Z Test for the Mean, sigma known.

For the example, select **PHStat → One-Sample Tests → Z Test for the Mean, sigma known**. In the procedure's dialog box (shown below):

1. Enter **368** as the **Null Hypothesis**.
2. Enter **0.05** as the **Level of Significance**.
3. Enter **15** as the **Population Standard Deviation**.
4. Click **Sample Statistics Known** and enter **25** as the **Sample Size** and **372.5** as the **Sample Mean**.
5. Click **Two-Tail Test**.
6. Enter a **Title** and click **OK**.



When using unsummarized data, click **Sample Statistics Unknown** in step 4 and enter the cell range of the unsummarized data as the **Sample Cell Range**.

In-Depth Excel Use the COMPUTE worksheet of the Z Mean workbook as a template.

The worksheet already contains the data for the example. For other problems, change the null hypothesis, level of significance, population standard deviation, sample size, and sample mean values in cells B4 through B8 as necessary.

Read the SHORT TAKES for Chapter 9 for an explanation of the formulas found in the COMPUTE worksheet. If you use an Excel version older than Excel 2010, use the COMPUTE_Older worksheet.

EG9.2 t TEST of HYPOTHESIS for the MEAN (σ UNKNOWN)

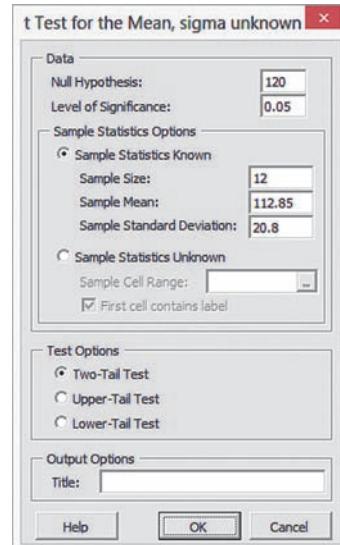
Key Technique Use the **T.INV.2T(level of significance, degrees of freedom)** function to compute the lower and upper critical values and use **T.DIST.2T(absolute value of the *t* test statistic, degrees of freedom)** to compute the *p*-value. Use an **IF** function (see Appendix Section F.4) to determine whether to display a rejection or nonrejection message.

Example Perform the Figure 9.7 two-tail *t* test for the mean for the sales invoices example shown on page 324.

PHStat Use t Test for the Mean, sigma unknown.

For the example, select **PHStat → One-Sample Tests → t Test for the Mean, sigma unknown**. In the procedure's dialog box (shown below):

1. Enter **120** as the **Null Hypothesis**.
2. Enter **0.05** as the **Level of Significance**.
3. Click **Sample Statistics Known** and enter **12** as the **Sample Size**, **112.85** as the **Sample Mean**, and **20.8** as the **Sample Standard Deviation**.
4. Click **Two-Tail Test**.
5. Enter a **Title** and click **OK**.



When using unsummarized data, click **Sample Statistics Unknown** in step 3 and enter the cell range of the unsummarized data as the **Sample Cell Range**.

In-Depth Excel Use the **COMPUTE worksheet** of the **T mean workbook**, as a template.

The worksheet already contains the data for the example. For other problems, change the values in cells B4 through B8 as necessary.

Read the **SHORT TAKES** for Chapter 9 for an explanation of the formulas found in the COMPUTE worksheet. If you use an Excel version older than Excel 2010, use the COMPUTE_OLDER worksheet.

EG9.3 ONE-TAIL TESTS

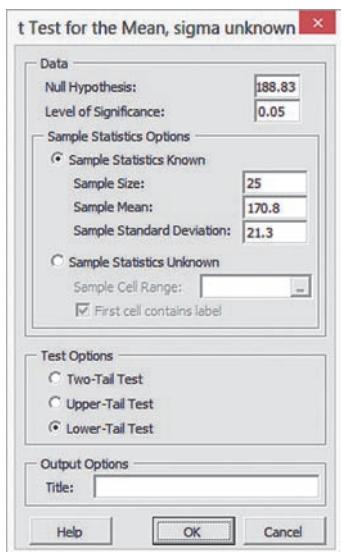
Key Technique Use the functions discussed in Section EG9.1 and EG9.2 to perform one-tail tests. For the *t* test of the mean, use **T.DIST.RT(*absolute value of the t test statistic, degrees of freedom*)** to help compute *p*-values. (See Appendix Section F.4.)

Example Perform the Figure 9.11 lower-tail *t* test for the mean for the drive-through time study example shown on page 330.

PHStat Click either **Lower-Tail Test** or **Upper-Tail Test** in the procedure dialog boxes discussed in Sections EG9.1 and EG9.2 to perform a one-tail test.

For the example, select **PHStat → One-Sample Tests → t Test for the Mean, sigma unknown**. In the procedure's dialog box (shown below):

1. Enter **188.83** as the **Null Hypothesis**.
2. Enter **0.05** as the **Level of Significance**.
3. Click **Sample Statistics Known** and enter **25** as the **Sample Size**, **170.8** as the **Sample Mean**, and **21.3** as the **Sample Standard Deviation**.
4. Click **Lower-Tail Test**.
5. Enter a **Title** and click **OK**.



In-Depth Excel Use the **COMPUTE_LOWER worksheet** or the **COMPUTE_UPPER worksheet** of the **Z Mean workbook** or the **T mean workbook** as templates.

For the example, open to the **COMPUTE_LOWER worksheet** of the **T mean workbook**.

Read the **SHORT TAKES** for Chapter 9 for an explanation of the formulas found in the worksheets. If you use an Excel version older than Excel 2010, use the **COMPUTE_OLDER worksheet**.

EG9.4 Z TEST of HYPOTHESIS for the PROPORTION

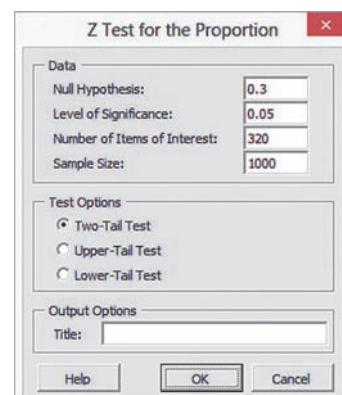
Key Technique Use the **NORM.S.INV** function to compute the lower and upper critical values and use **NORM.S.DIST(*absolute value of the Z test statistic, True*)** as part of a formula to compute the *p*-value. Use an **IF** function (see Appendix Section F.4) to determine whether to display a rejection or nonrejection message.

Example Perform the Figure 9.14 two-tail *Z* test for the proportion of all adults who were unable to stop thinking about work while on vacation shown on page 334.

PHStat Use **Z Test for the Proportion**.

For the example, select **PHStat → One-Sample Tests → Z Test for the Proportion**. In the procedure's dialog box (shown below):

1. Enter **0.3** as the **Null Hypothesis**.
2. Enter **0.05** as the **Level of Significance**.
3. Enter **320** as the **Number of Items of Interest**.
4. Enter **1000** as the **Sample Size**.
5. Click **Two-Tail Test**.
6. Enter a **Title** and click **OK**.



In-Depth Excel Use the **COMPUTE worksheet** of the **Z Proportion workbook** as a template.

The worksheet already contains the data for the example. For other problems, change the null hypothesis, level of significance, population standard deviation, sample size, and sample mean values in cells B4 through B7 as necessary.

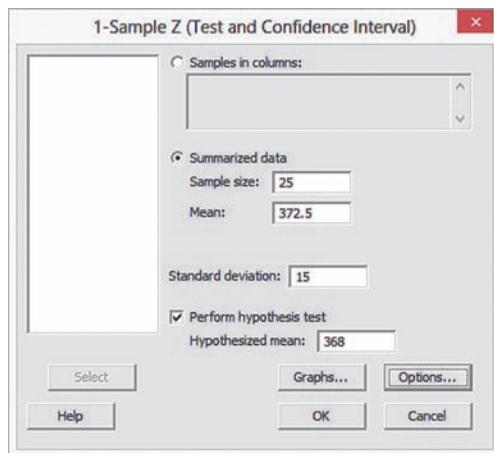
Read the **SHORT TAKES** for Chapter 9 for an explanation of the formulas found in the COMPUTE worksheet. Use the **COMPUTE_LOWER** or **COMPUTE_UPPER** worksheets as templates for performing one-tail tests. If you use an Excel version older than Excel 2010, use the **COMPUTE_OLDER** worksheet.

CHAPTER 9 MINITAB GUIDE

MG9.1 FUNDAMENTALS of HYPOTHESIS-TESTING METHODOLOGY

Use **1-Sample Z** to perform the *Z* test for the mean when σ is known. For example, to perform the two-tail *Z* test for the Figure 9.5 cereal-filling example on page 318, select **Stat → Basic Statistics → 1-Sample Z**. In the “1-Sample Z (Test and Confidence Interval)” dialog box (shown below):

1. Click **Summarized data**.
2. Enter **25** in the **Sample size** box and **372.5** in the **Mean** box.
3. Enter **15** in the **Standard deviation** box.
4. Check **Perform hypothesis test** and enter **368** in the **Hypothesized mean** box.
5. Click **Options**.



In the 1-Sample Z - Options dialog box:

6. Enter **95.0** in the **Confidence level** box.
7. Select **not equal** from the **Alternative** drop-down list.
8. Click **OK**.
9. Back in the original dialog box, click **OK**.

When using unsummarized data, open the worksheet that contains the data and replace steps 1 and 2 with these steps:

1. Click **Samples in columns**.
2. Enter the name of the column containing the unsummarized data in the **Samples in column** box.

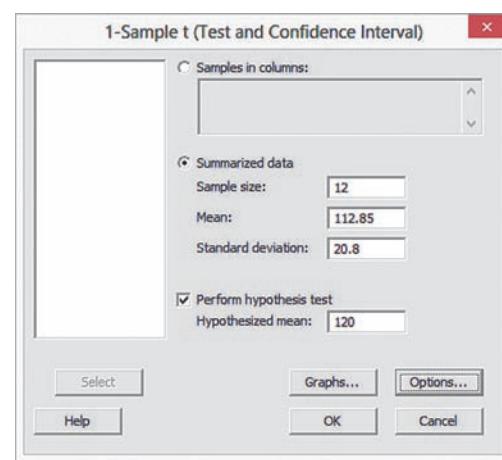
MG9.2 t TEST of HYPOTHESIS for the MEAN (σ UNKNOWN)

Use **1-Sample t** to perform the *t* test for the mean when σ is unknown.

For example, to perform the *t* test for the Figure 9.7 sales invoice example on page 324, select **Stat → Basic Statistics → 1-Sample t**. In the 1-Sample t (Test and Confidence Interval) dialog box (shown in the next column):

1. Click **Summarized data**.

2. Enter **12** in the **Sample size** box, **112.85** in the **Mean** box, and **20.8** in the **Standard deviation** box.
 3. Check **Perform hypothesis test** and enter **120** in the **Hypothesized mean** box.
 4. Click **Options**.
- In the 1-Sample t - Options dialog box:
5. Enter **95.0** in the **Confidence level** box.
 6. Select **not equal** from the **Alternative** drop-down list.
 7. Click **OK**.
 8. Back in the original dialog box, click **OK**.



When using unsummarized data, open the worksheet that contains the data and replace steps 1 and 2 with these steps:

1. Click **Samples in columns**.
 2. Enter the name of the column containing the unsummarized data in the **Samples in column** box.
- To create a boxplot of the unsummarized data, replace step 8 with the following steps 8 through 10:
8. Back in the original dialog box, click **Graphs**.
 9. In the 1-Sample t - Graphs dialog box, check **Boxplot of data** and then click **OK**.
 10. Back in the original dialog box, click **OK**.

MG9.3 ONE-TAIL TESTS

To perform a one-tail test for **1-Sample Z**, select **less than** or **greater than** from the drop-down list in step 7 of the Section MG9.1 instructions.

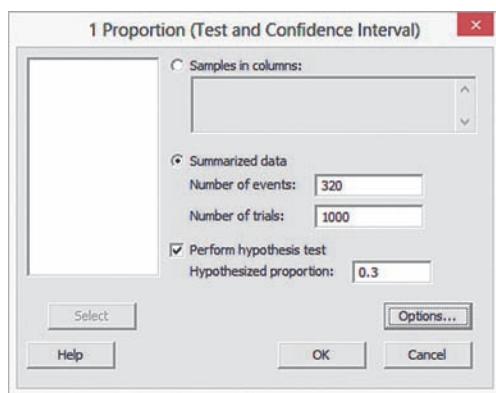
To perform a one-tail test for **1-Sample t**, select **less than** or **greater than** from the drop-down list in step 6 of the Section MG9.2 instructions.

MG9.4 Z TEST of HYPOTHESIS for the PROPORTION

Use 1 Proportion.

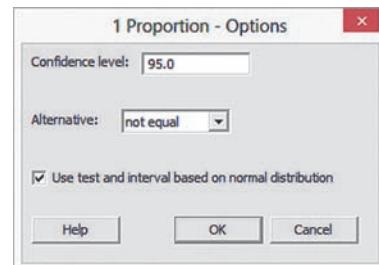
For example, to perform the Figure 9.14 Z test for the proportion of all adults who were unable to stop thinking about work while on vacation on page 334, select **Stat → Basic Statistics → 1 Proportion**. In the 1 Proportion (Test and Confidence Interval) dialog box (shown below):

1. Click **Summarized data**.
2. Enter **320** in the **Number of events** box and **1000** in the **Number of trials** box.
3. Check **Perform hypothesis test** and enter **0.3** in the **Hypothesized proportion** box.
4. Click **Options**.



In the 1-Proportion - Options dialog box (shown below):

5. Enter **95.0** in the **Confidence level** box.
6. Select **not equal** from the **Alternative** drop-down list.
7. Check **Use test and interval based on normal distribution**.
8. Click **OK**.



9. Back in the original dialog box, click **OK**.

When using unsummarized data, open the worksheet that contains the data and replace steps 1 and 2 with these steps:

1. Click **Samples in columns**.
2. Enter the name of the column containing the unsummarized in the **Samples in column** box.

To perform a one-tail test, select **less than** or **greater than** from the drop-down list in step 6.

CHAPTER 10

Two-Sample Tests

CONTENTS

10.1 Comparing the Means of Two Independent Populations

Do People Really Do This?

10.2 Comparing the Means of Two Related Populations

10.3 Comparing the Proportions of Two Independent Populations

10.4 *F* Test for the Ratio of Two Variances

USING STATISTICS: For North Fork, Are There Different Means to the Ends? Revisited

CHAPTER 10 EXCEL GUIDE

CHAPTER 10 MINITAB GUIDE

OBJECTIVES

To compare the means of two independent populations

To compare the means of two related populations

To compare the proportions of two independent populations

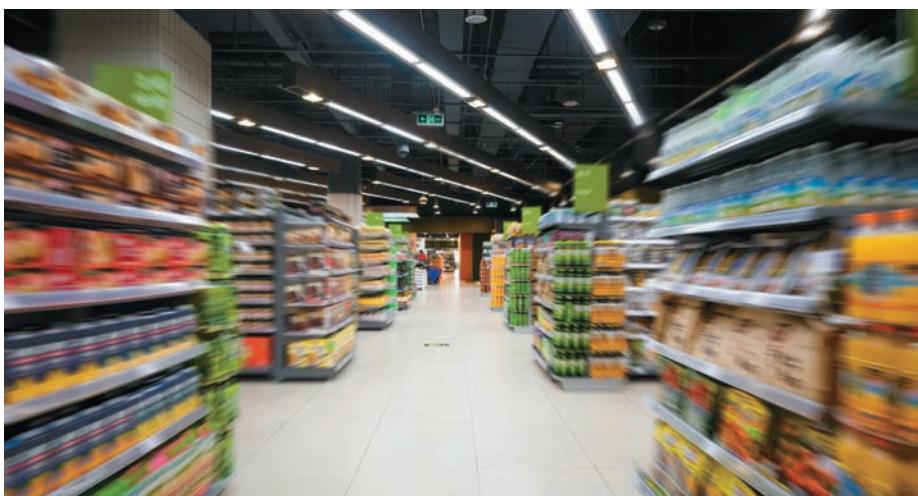
To compare the variances of two independent populations

USING STATISTICS

For North Fork, Are There Different Means to the Ends?

To what extent does the location of products affect sales in a supermarket? As a North Fork Beverages sales manager, you are negotiating with the management of FoodPlace Supermarkets for the location of displays for the new HandMade Real Citrus Cola. FoodPlace Supermarkets has offered you two different end-aisle display areas to feature your new cola: one near the produce department and the other at the front of the aisle that contains other beverage products. These ends of aisle, or end-caps, have different costs, and you would like to compare the effectiveness of the produce end-cap to the beverage end-cap.

To test the comparative effectiveness of the two end-caps, FoodPlace agrees to a pilot study. You will be able to select 20 stores from the supermarket chain that experience similar storewide sales volumes. You then randomly assign 10 of the 20 stores to sample 1 and 10 other stores to sample 2. In the sample 1 stores, you will place the new cola in the beverage end-cap, while in the sample 2 stores you will place the new cola in the produce end-cap. At the end of one week, the sales of the new cola will be recorded. How can you determine whether the sales of the new cola using beverage end-caps are different from the sales of the new cola using produce end-caps? How can you decide if the variability in new cola sales from store to store is different for the two types of displays? How could you use the answers to these questions to improve sales of your new HandMade Real Citrus Cola?



Fotolia

In Chapter 9, you learned several hypothesis-testing procedures commonly used to test a single sample of data selected from a single population. In this chapter, you learn how to extend hypothesis testing to **two-sample tests** that compare statistics from samples selected from *two* populations. In the North Fork Beverages scenario one such test would be “Are the mean weekly sales of the new cola when using the beverage end-cap location (one population) different from the mean weekly sales of the new cola when using the produce end-cap location (a second population)?”

10.1 Comparing the Means of Two Independent Populations

In Sections 8.1 and 9.1, you learned that in almost all cases, you would not know the standard deviation of the population under study. Likewise, when you take a random sample from each of two independent populations, you almost always do not know the standard deviation of either population. In addition, when using a two-sample test that compares the means of samples selected from two populations, you must establish whether the assumption that the variances in the two populations are equal holds. The statistical method used to test whether the means of each population are different depends on whether the assumption holds or not.

Pooled-Variance *t* Test for the Difference Between Two Means

If you assume that the random samples are independently selected from two populations and that the populations are normally distributed and have equal variances, you can use a **pooled-variance *t* test** to determine whether there is a significant difference between the means. If the populations do not differ greatly from a normal distribution, you can still use the pooled-variance *t* test, especially if the sample sizes are large enough (typically ≥ 30 for each sample).

Using subscripts to distinguish between the population mean of the first population, μ_1 , and the population mean of the second population, μ_2 , the null hypothesis of no difference in the means of two independent populations can be stated as

$$H_0: \mu_1 = \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 = 0$$

and the alternative hypothesis, that the means are different, can be stated as

$$H_1: \mu_1 \neq \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 \neq 0$$

To test the null hypothesis, you use the pooled-variance *t* test statistic t_{STAT} shown in Equation (10.1). The pooled-variance *t* test gets its name from the fact that the test statistic pools, or combines, the two sample variances S_1^2 and S_2^2 to compute S_p^2 , the best estimate of the variance common to both populations, under the assumption that the two population variances are equal.¹

Student Tip

Whichever population is defined as population 1 in the null and alternative hypotheses must be defined as population 1 in Equation (10.1). Whichever population is defined as population 2 in the null and alternative hypotheses must be defined as population 2 in Equation (10.1).

POOLED-VARIANCE *t* TEST FOR THE DIFFERENCE BETWEEN TWO MEANS

$$t_{STAT} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (10.1)$$

where $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$

and S_p^2 = pooled variance

\bar{X}_1 = mean of the sample taken from population 1

¹When the two sample sizes are equal (i.e., $n_1 = n_2$), the equation for the pooled variance can be simplified to

$$S_p^2 = \frac{S_1^2 + S_2^2}{2}$$

S_1^2 = variance of the sample taken from population 1

n_1 = size of the sample taken from population 1

\bar{X}_2 = mean of the sample taken from population 2

S_2^2 = variance of the sample taken from population 2

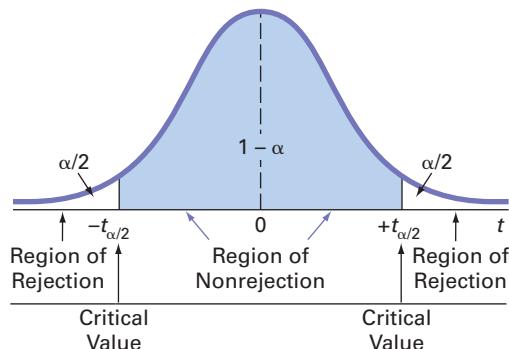
n_2 = size of the sample taken from population 2

The t_{STAT} test statistic follows a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

For a given level of significance, α , in a two-tail test, you reject the null hypothesis if the computed t_{STAT} test statistic is greater than the upper-tail critical value from the t distribution or if the computed t_{STAT} test statistic is less than the lower-tail critical value from the t distribution. Figure 10.1 displays the regions of rejection.

FIGURE 10.1

Regions of rejection and nonrejection for the pooled-variance t test for the difference between the means (two-tail test)



Student Tip

When *lower or less than* is used in an example, you have a lower-tail test. When *upper or more than* is used in an example, you have an upper-tail test. When *different or the same as* is used in an example, you have a two-tail test.

In a one-tail test in which the rejection region is in the lower tail, you reject the null hypothesis if the computed t_{STAT} test statistic is less than the lower-tail critical value from the t distribution. In a one-tail test in which the rejection region is in the upper tail, you reject the null hypothesis if the computed t_{STAT} test statistic is greater than the upper-tail critical value from the t distribution.

To demonstrate the pooled-variance t test, return to the North Fork Beverages scenario on page 347. Using the DCOVA problem-solving approach, you define the business objective as determining whether there is a difference in the mean weekly sales of the new cola when using the beverage end-cap location and when using the produce end-cap location. There are two populations of interest. The first population is the set of all possible weekly sales of the new cola if all the FoodPlace Supermarkets used the beverage end-cap location. The second population is the set of all possible weekly sales of the new cola if all the FoodPlace Supermarkets used the produce end-cap location. You collect the data from a sample of 10 FoodPlace Supermarkets that have been assigned a beverage end-cap location and another sample of 10 FoodPlace Supermarkets that have been assigned a produce end-cap location. You organize and store the results in **Cola**. Table 10.1 contains the new cola sales (in number of cases) for the two samples.

TABLE 10.1

Comparing New Cola Weekly Sales from Two Different End-Cap Locations (in number of cases)

DISPLAY LOCATION									
Beverage End-Cap					Produce End-Cap				
22	34	52	62	30	52	71	76	54	67
40	64	84	56	59	83	66	90	77	84

The null and alternative hypotheses are

$$H_0: \mu_1 = \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 \neq \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 \neq 0$$

Assuming that the samples are from normal populations having equal variances, you can use the pooled-variance t test. The t_{STAT} test statistic follows a t distribution with

$10 + 10 - 2 = 18$ degrees of freedom. Using an $\alpha = 0.05$ level of significance, you divide the rejection region into the two tails for this two-tail test (i.e., two equal parts of 0.025 each). Table E.3 shows that the critical values for this two-tail test are $+2.1009$ and -2.1009 . As shown in Figure 10.2, the decision rule is

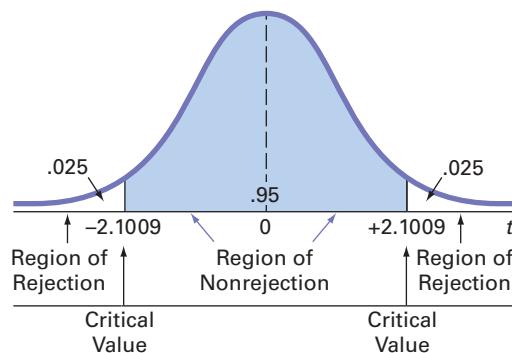
Reject H_0 if $t_{STAT} > +2.1009$

or if $t_{STAT} < -2.1009$;

otherwise, do not reject H_0 .

FIGURE 10.2

Two-tail test of hypothesis for the difference between the means at the 0.05 level of significance with 18 degrees of freedom



From Figure 10.3, the computed t_{STAT} test statistic for this test is -3.0446 and the p -value is 0.0070.

FIGURE 10.3

Excel and Minitab pooled-variance t test results for the two end-cap locations data

Pooled-Variance t Test for Differences in Two Means		B	Two-Sample T-Test and CI: Beverage, Produce				
(assumes equal population variances)			Two-sample T for Beverage vs Produce				
Data			N	Mean	StDev	SE Mean	
4	Hypothesized Difference	0	Beverage	10	50.3	18.7	5.9
5	Level of Significance	0.05	Produce	10	72.0	12.5	4.0
Population 1 Sample			Difference = mu (Beverage) - mu (Produce)				
7	Sample Size	10	=COUNT(DATACOPY!\$A:\$A)	Estimate for difference: -21.70			
8	Sample Mean	50.3	=AVERAGE(DATACOPY!\$A:\$A)	95% CI for difference: (-36.67, -6.73)			
9	Sample Standard Deviation	18.7264	=STDEV.S(DATACOPY!\$A:\$A)	T-Test of difference = 0 (vs not =):			
Population 2 Sample			T-Value = -3.04 P-Value = 0.007 DF = 18				
11	Sample Size	10	=COUNT(DATACOPY!\$B:\$B)	Both use Pooled StDev = 15.9376			
12	Sample Mean	72	=AVERAGE(DATACOPY!\$B:\$B)				
13	Sample Standard Deviation	12.5433	=STDEV.S(DATACOPY!\$B:\$B)				
Intermediate Calculations							
16	Population 1 Sample Degrees of Freedom	9	=B7 - 1				
17	Population 2 Sample Degrees of Freedom	9	=B11 - 1				
18	Total Degrees of Freedom	18	=B16 + B17				
19	Pooled Variance	254.0056	=((B16 * B9^2) + (B17 * B13^2)) / B18				
20	Standard Error	7.1275	=SQRT(B19 * (1/B7 + 1/B11))				
21	Difference in Sample Means	-21.7	=B8 - B12				
22	t Test Statistic	-3.0446	=(B21 - B4) / B20				
Two-Tail Test							
25	Lower Critical Value	-2.1009	=T.INV.2T(B5, B18)				
26	Upper Critical Value	2.1009	=T.INV.2T(B5, B18)				
27	p-Value	0.0070	=T.DIST.2T(ABS(B22), B18)				
28	Reject the null hypothesis		=IF(B27 < B5, "Reject the null hypothesis", "Do not reject the null hypothesis")				

Using Equation (10.1) on page 348 and the descriptive statistics provided in Figure 10.3,

$$t_{STAT} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{9(18.7264)^2 + 9(12.5433)^2}{9 + 9} = 254.0056 \end{aligned}$$

Therefore,

$$t_{STAT} = \frac{(50.3 - 72.0) - 0.0}{\sqrt{254.0056 \left(\frac{1}{10} + \frac{1}{10} \right)}} = \frac{-21.7}{\sqrt{50.801}} = -3.0446$$

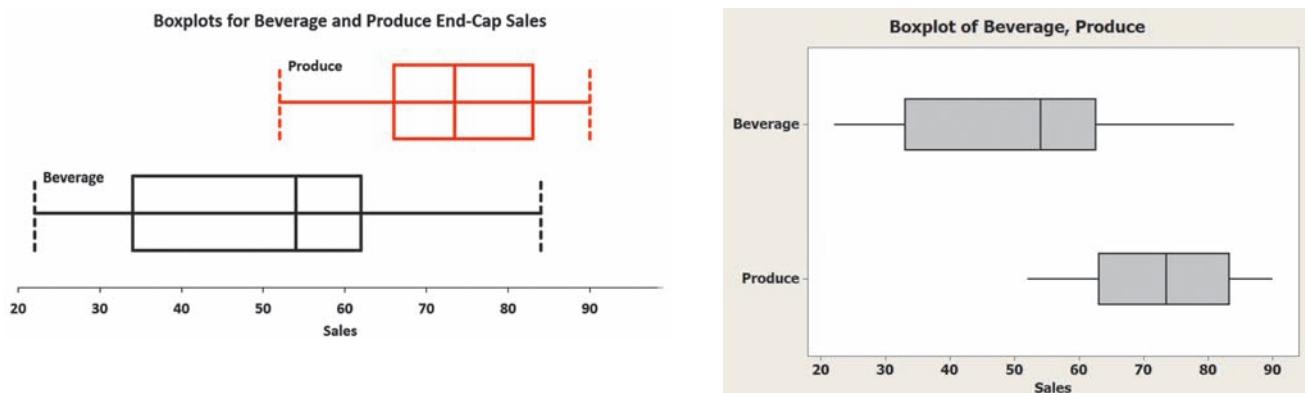
You reject the null hypothesis because $t_{STAT} = -3.0446 < -2.1009$ and the p -value is 0.0070. In other words, the probability that $t_{STAT} > 3.0446$ or $t_{STAT} < -3.0446$ is equal to 0.0070. This p -value indicates that if the population means are equal, the probability of observing a difference this large or larger in the two sample means is only 0.0070. Because the p -value is less than $\alpha = 0.05$, there is sufficient evidence to reject the null hypothesis. You can conclude that the mean sales are different for the beverage end-cap and produce end-cap locations. Because the t_{STAT} statistic is negative, you can conclude that the mean sales are lower for the beverage end-cap location (and, therefore, higher for the produce end-cap location).

In testing for the difference between the means, you assume that the populations are normally distributed, with equal variances. For situations in which the two populations have equal variances, the pooled-variance t test is **robust** (i.e., not sensitive) to moderate departures from the assumption of normality, provided that the sample sizes are large. In such situations, you can use the pooled-variance t test without serious effects on its power. However, if you cannot assume that both populations are normally distributed, you have two choices. You can use a nonparametric procedure, such as the Wilcoxon rank sum test (see Section 12.4), that does not depend on the assumption of normality for the two populations, or you can use a normalizing transformation (see reference 6) on each of the outcomes and then use the pooled-variance t test.

To check the assumption of normality in each of the two populations, you can construct a boxplot of the sales for the two display locations shown in Figure 10.4. For these two small samples, there appears to be only moderate departure from normality, so the assumption of normality needed for the t test is not seriously violated.

FIGURE 10.4

Excel and Minitab boxplots for beverage and produce end-cap sales



Example 10.1 provides another application of the pooled-variance t test.

EXAMPLE 10.1

Testing for the Difference in the Mean Delivery Times

You and some friends have decided to test the validity of an advertisement by a local pizza restaurant, which says it delivers to the dormitories faster than a local branch of a national chain. Both the local pizza restaurant and national chain are located across the street from your college campus. You define the variable of interest as the delivery time, in minutes, from the time the pizza is ordered to when it is delivered. You collect the data by ordering 10 pizzas from the local pizza restaurant and 10 pizzas from the national chain at different times. You organize and store the data in **PizzaTime**. Table 10.2 shows the delivery times.

(continued)

TABLE 10.2

Delivery Times (in minutes) for a Local Pizza Restaurant and a National Pizza Chain

	Local		Chain	
	16.8	18.1	22.0	19.5
	11.7	14.1	15.2	17.0
	15.6	21.8	18.7	19.5
	16.7	13.9	15.6	16.5
	17.5	20.8	20.8	24.0

At the 0.05 level of significance, is there evidence that the mean delivery time for the local pizza restaurant is less than the mean delivery time for the national pizza chain?

SOLUTION Because you want to know whether the mean is *lower* for the local pizza restaurant than for the national pizza chain, you have a one-tail test with the following null and alternative hypotheses:

$H_0: \mu_1 \geq \mu_2$ (The mean delivery time for the local pizza restaurant is equal to or greater than the mean delivery time for the national pizza chain.)

$H_1: \mu_1 < \mu_2$ (The mean delivery time for the local pizza restaurant is less than the mean delivery time for the national pizza chain.)

Figure 10.5 displays the results for the pooled-variance t test for these data.

FIGURE 10.5

Excel and Minitab pooled-variance t test results for the pizza delivery time data

A		B
1 Pooled-Variance t Test for Differences in Two Means		
2 (assumes equal population variances)		
3 Data		
4 Hypothesized Difference		0
5 Level of Significance		0.05
6 Population 1 Sample		
7 Sample Size		10
8 Sample Mean		16.7
9 Sample Standard Deviation		3.0955
10 Population 2 Sample		
11 Sample Size		10
12 Sample Mean		18.88
13 Sample Standard Deviation		2.8662
14		
15 Intermediate Calculations		
16 Population 1 Sample Degrees of Freedom		9
17 Population 2 Sample Degrees of Freedom		9
18 Total Degrees of Freedom		18
19 Pooled Variance		8.8986
20 Standard Error		1.3341
21 Difference in Sample Means		-2.18
22 t Test Statistic		-1.6341
23		
24 Lower-Tail Test		
25 Lower Critical Value		-1.7341
26 p-Value		0.0598
27 Do not reject the null hypothesis		

Two-Sample T-Test and CI: Local, Chain				
Two-sample T for Local vs Chain				
	N	Mean	StDev	SE Mean
Local	10	16.70	3.10	0.98
Chain	10	18.88	2.87	0.91
Difference = mu (Local) - mu (Chain)				
Estimate for difference:		-2.18		
95% CI for difference:		(-4.98, 0.62)		
T-Test of difference = 0 (vs not =): T-Value = -1.63 P-Value = 0.120 DF = 18				
Both use Pooled StDev = 2.9831				

To illustrate the computations, using Equation (10.1) on page 348,

$$t_{STAT} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{9(3.0955)^2 + 9(2.8662)^2}{9 + 9} = 8.8986 \end{aligned}$$

Therefore,

$$t_{STAT} = \frac{(16.7 - 18.88) - 0.0}{\sqrt{8.8986\left(\frac{1}{10} + \frac{1}{10}\right)}} = \frac{-2.18}{\sqrt{1.7797}} = -1.6341$$

You do not reject the null hypothesis because $t_{STAT} = -1.6341 > -1.7341$. The p -value (as computed in Figure 10.5) is 0.0598. This p -value indicates that the probability that $t_{STAT} < -1.6341$ is equal to 0.0598. In other words, if the population means are equal, the probability that the sample mean delivery time for the local pizza restaurant is at least 2.18 minutes faster than the national chain is 0.0598. Because the p -value is greater than $\alpha = 0.05$, there is insufficient evidence to reject the null hypothesis. Based on these results, there is insufficient evidence for the local pizza restaurant to make the advertising claim that it has a faster delivery time.

Confidence Interval Estimate for the Difference Between Two Means

Instead of, or in addition to, testing for the difference between the means of two independent populations, you can use Equation (10.2) to develop a confidence interval estimate of the difference in the means.

CONFIDENCE INTERVAL ESTIMATE FOR THE DIFFERENCE BETWEEN THE MEANS OF TWO INDEPENDENT POPULATIONS

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

or

$$(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (10.2)$$

where $t_{\alpha/2}$ is the critical value of the t distribution, with $n_1 + n_2 - 2$ degrees of freedom, for an area of $\alpha/2$ in the upper tail.

For the sample statistics pertaining to the two end-cap locations reported in Figure 10.3 on page 350, using 95% confidence, and Equation (10.2),

$$\begin{aligned} \bar{X}_1 &= 50.3, n_1 = 10, \bar{X}_2 = 72.0, n_2 = 10, S_p^2 = 254.0056, \text{ and with } 10 + 10 - 2 \\ &= 18 \text{ degrees of freedom, } t_{0.025} = 2.1009 \\ (50.3 - 72.0) &\pm (2.1009) \sqrt{254.0056 \left(\frac{1}{10} + \frac{1}{10} \right)} \\ &-21.7 \pm (2.1009)(7.1275) \\ &-21.7 \pm 14.97 \\ -36.67 &\leq \mu_1 - \mu_2 \leq -6.73 \end{aligned}$$

Therefore, you are 95% confident that the difference in mean sales between the beverage and produce end-cap locations is between -36.67 cases of cola and -6.73 cases of cola. In other words, you can estimate, with 95% confidence, that the produce end-cap location sells, on average, 6.73 to 36.67 cases more than the beverage end-cap location. From a hypothesis-testing perspective, using a two-tail test at the 0.05 level of significance, because the interval does not include zero, you reject the null hypothesis of no difference between the means of the two populations.

t Test for the Difference Between Two Means, Assuming Unequal Variances

If you can assume that the two independent populations are normally distributed but cannot assume that they have equal variances, you cannot pool the two sample variances into the common estimate S_p^2 and therefore cannot use the pooled-variance t test. Instead, you use the **separate-variance t test** developed by Satterthwaite (see reference 5). Equation (10.3) defines the test statistic for the separate-variance t test.

SEPARATE-VARIANCE t TEST FOR THE DIFFERENCE BETWEEN TWO MEANS

$$t_{STAT} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (10.3)$$

where

- \bar{X}_1 = mean of the sample taken from population 1
- S_1^2 = variance of the sample taken from population 1
- n_1 = size of the sample taken from population 1
- \bar{X}_2 = mean of the sample taken from population 2
- S_2^2 = variance of the sample taken from population 2
- n_2 = size of the sample taken from population 2

The separate-variance t test statistic approximately follows a t distribution with degrees of freedom V equal to the integer portion of the following computation shown in Equation (10.4).

COMPUTING DEGREES OF FREEDOM IN THE SEPARATE-VARIANCE t TEST

$$V = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{\left(\frac{S_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2} \right)^2}{n_2 - 1}} \quad (10.4)$$

For a given level of significance α , you reject the null hypothesis if the computed t test statistic is greater than the upper-tail critical value $t_{a/2}$ from the t distribution with V degrees of freedom or if the computed test statistic is less than the lower-tail critical value $-t_{a/2}$ from the t distribution with V degrees of freedom. Thus, the decision rule is

Reject H_0 if $t > t_{a/2}$
or if $t < -t_{a/2}$;
otherwise, do not reject H_0 .

Return to the North Fork Beverages scenario concerning the two end-cap display locations. Using Equation (10.4), the separate-variance t test statistic t_{STAT} is approximated

by a t distribution with $V = 15$ degrees of freedom, the integer portion of the following computation:

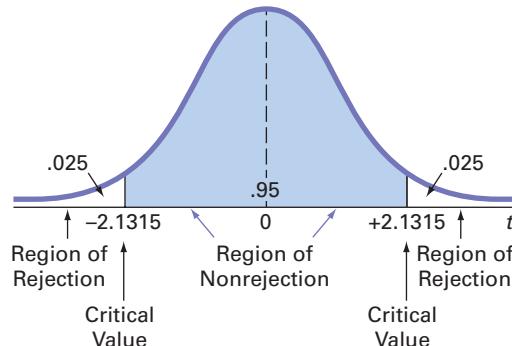
$$\begin{aligned} V &= \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} \\ &= \frac{\left(\frac{350.6778}{10} + \frac{157.3333}{10}\right)^2}{\left(\frac{350.6778}{10}\right)^2 + \left(\frac{157.3333}{10}\right)^2} = 15.72 \end{aligned}$$

Using $\alpha = 0.05$, the upper and lower critical values for this two-tail test found in Table E.3 are $+2.1315$ and -2.1315 . As depicted in Figure 10.6, the decision rule is

Reject H_0 if $t_{STAT} > +2.1315$
or if $t_{STAT} < -2.1315$;
otherwise, do not reject H_0 .

FIGURE 10.6

Two-tail test of hypothesis for the difference between the means at the 0.05 level of significance with 15 degrees of freedom



Using Equation (10.3) on page 354 and the descriptive statistics provided in Figure 10.3,

$$\begin{aligned} t_{STAT} &= \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \\ &= \frac{50.3 - 72}{\sqrt{\left(\frac{350.6778}{10} + \frac{157.3333}{10}\right)}} = \frac{-21.7}{\sqrt{50.801}} = -3.04 \end{aligned}$$

Using a 0.05 level of significance, you reject the null hypothesis because $t = -3.04 < -2.1315$.

Figure 10.7 on page 356 displays the separate-variance t test results for the end-cap display location data. Observe that the test statistic $t_{STAT} = -3.0446$ and the p -value is $0.0082 < 0.05$. Thus, the results for the separate-variance t test are nearly the same as those of the pooled-variance t test. The assumption of equality of population variances had no appreciable effect on the results. Sometimes, however, the results from the pooled-variance and separate-variance t tests conflict because the assumption of equal variances is violated. Therefore, it is important that you evaluate the assumptions and use those results as a guide in selecting a test procedure.

FIGURE 10.7

Excel and Minitab separate-variance t test results for the sales data for the two end-caps

A	B
1 Separate-Variances t Test	
2 (assumes unequal population variances)	
3 Data	
4 Hypothesized Difference	0
5 Level of Significance	0.05
6 Population 1 Sample	
7 Sample Size	10 =COUNT(DATACOPY!\$A:\$A)
8 Sample Mean	50.3 =AVERAGE(DATACOPY!\$A:\$A)
9 Sample Standard Deviation	18.7264 =STDEV.S(DATACOPY!\$A:\$A)
10 Population 2 Sample	
11 Sample Size	10 =COUNT(DATACOPY!\$B:\$B)
12 Sample Mean	72 =AVERAGE(DATACOPY!\$B:\$B)
13 Sample Standard Deviation	12.5433 =STDEV.S(DATACOPY!\$B:\$B)
14	
15 Intermediate Calculations	
16 Pop. 1 Sample Variance	350.6778 =B9^2
17 Pop. 2 Sample Variance	157.3333 =B13^2
18 Pop. 1 Sample Var./Sample Size	35.0678 =B16/B7
19 Pop. 2 Sample Var./Sample Size	15.7333 =B17/B11
20 Numerator of Degrees of Freedom	2580.7529 =(B18 + B19)^2
21 Denominator of Degrees of Freedom	164.1430 =(B18^2)/(B7 - 1) + (B19^2)/(B11 - 1)
22 Total Degrees of Freedom	15.7226 =B20/B21
23 Degrees of Freedom	15 =INT(B22)
24 Separate Variance Denominator	7.1275 =SQRT(B18 + B19)
25 Difference in Sample Means	-21.7 =B8 - B12
26 t Test Statistic	-3.0446 =(B25 - B4)/B24
27	
28 Two-Tail Test	
29 Lower Critical Value	-2.1314 =-(T.INV.2T(B5, B23))
30 Upper Critical Value	2.1314 =T.INV.2T(B5, B23)
31 p -Value	0.0082 =T.DIST.2T(ABS(B26), B23) - B4
32 Reject the null hypothesis	=IF(B31 < B5, "Reject the null hypothesis", "Do not reject the null hypothesis")

Two-Sample T-Test and CI: Beverage, Produce**Two-sample T for Beverage vs Produce**

	N	Mean	StDev	SE Mean
Beverage	10	50.3	18.7	5.9
Produce	10	72.0	12.5	4.0

Difference = mu (Beverage) - mu (Produce)
Estimate for difference: -21.70
95% CI for difference: (-36.89, -6.51)
T-Test of difference = 0 (vs not =):
T-Value = -3.04 P-Value = 0.008 DF = 15

In Section 10.4, the F test for the ratio of two variances is used to determine whether there is evidence of a difference in the two population variances. The results of that test can help you decide which of the t tests—pooled-variance or separate-variance—is more appropriate.

Do People Really Do This?

You may have heard opinions that lead you to wonder if decision-makers really use confirmatory methods, such as hypothesis testing, in this emerging era of big data. The following real case study, contributed by a former student of a colleague of the authors, reveals a role that confirmatory methods still play in business as well as answering another question: “Do businesses really monitor their customer service calls for quality assurance purposes as they sometime claim?”

In her first full-time job at a financial services company, a student was asked to improve a training program for new hires at a call center that handled customer questions about outstanding loans. For feedback and evaluation, she planned to randomly select phone calls received by each new employee and rate the employee on 10 aspects of the call, including whether the employee maintained a pleasant tone with the customer.

When she presented her plan to her boss for approval, her boss wanted proof that her new training program would improve customer service. The boss, quoting a famous statistician, said “In God we trust; all others must bring data.” Faced with this request, she called her business statistics professor. “Hello, Professor, you’ll never believe why I called. I work for a large company, and in the project I am currently working on, I have to put some of the statistics you taught us to work! Can you help?” Together they formulated this test:

- Randomly assign the 60 most recent hires to two training programs. Assign half to the preexisting training program and the other half to the new training program.
- At the end of the first month, compare the mean score for the 30 employees in the new training program against the mean score for

the 30 employees in the preexisting training program.

She listened as her professor explained, “What you are trying to show is that the mean score from the new training program is higher than the mean score from the current program. You can make the null hypothesis that the means are equal and see if you can reject it in favor of the alternative that the mean score from the new program is higher.”

“Or, as you used to say, ‘if the p -value is low, H_0 must go!’—yes, I do remember!” she replied. Her professor chuckled and added, “If you can reject H_0 you will have the evidence to present to your boss.” She thanked him for his help and got back to work, with the newfound confidence that she would be able to successfully apply the t test that compares the means of two independent populations.

Problems for Section 10.1

LEARNING THE BASICS

10.1 If you have samples of $n_1 = 12$ and $n_2 = 15$, in performing the pooled-variance t test, how many degrees of freedom do you have?

10.2 Assume that you have a sample of $n_1 = 8$, with the sample mean $\bar{X}_1 = 42$, and a sample standard deviation $S_1 = 4$, and you have an independent sample of $n_2 = 15$ from another population with a sample mean of $\bar{X}_2 = 34$ and a sample standard deviation $S_2 = 5$.

- What is the value of the pooled-variance t_{STAT} test statistic for testing $H_0: \mu_1 = \mu_2$?
- In finding the critical value, how many degrees of freedom are there?
- Using the level of significance $\alpha = 0.01$, what is the critical value for a one-tail test of the hypothesis $H_0: \mu_1 \leq \mu_2$ against the alternative, $H_1: \mu_1 > \mu_2$?
- What is your statistical decision?

10.3 What assumptions about the two populations are necessary in Problem 10.2?

10.4 Referring to Problem 10.2, construct a 95% confidence interval estimate of the population mean difference between μ_1 and μ_2 .

10.5 Referring to Problem 10.2, if $n_1 = 5$ and $n_2 = 4$, how many degrees of freedom do you have?

10.6 Referring to Problem 10.2, if $n_1 = 5$ and $n_2 = 4$, at the 0.01 level of significance, is there evidence that $\mu_1 > \mu_2$?

APPLYING THE CONCEPTS

10.7 When people make estimates, they are influenced by anchors to their estimates. A study was conducted in which students were asked to estimate the number of calories in a cheeseburger. One group was asked to do this after thinking about a calorie-laden cheesecake. A second group was asked to do this after thinking about an organic fruit salad. The mean number of calories estimated in a cheeseburger was 780 for the group that thought about the cheesecake and 1,041 for the group that thought about the organic fruit salad. (Data extracted from “Drilling Down, Sizing Up a Cheeseburger’s Caloric Heft,” *The New York Times*, October 4, 2010, p. B2.) Suppose that the study was based on a sample of 20 people who thought about the cheesecake first and 20 people who thought about the organic fruit salad first, and the standard deviation of the number of calories in the cheeseburger was 128 for the people who thought about the cheesecake first and 140 for the people who thought about the organic fruit salad first.

- State the null and alternative hypotheses if you want to determine whether the mean estimated number of calories in the cheeseburger is lower for the people who thought about the cheesecake first than for the people who thought about the organic fruit salad first.
- In the context of this study, what is the meaning of the Type I error?
- In the context of this study, what is the meaning of the Type II error?

- At the 0.01 level of significance, is there evidence that the mean estimated number of calories in the cheeseburger is lower for the people who thought about the cheesecake first than for the people who thought about the organic fruit salad first?

10.8 A recent study (data extracted from E. J. Boyland et al., “Food Choice and Overconsumption: Effect of a Premium Sports Celebrity Endorser,” *Journal of Pediatrics*, March 13, 2013, bit.ly/16NR4Bi) found that 51 children who watched a commercial for Walker Crisps (potato chips) featuring a long-standing sports celebrity endorser ate a mean of 36 grams of Walker Crisps as compared to a mean of 25 grams of Walker Crisps for 41 children who watched a commercial for an alternative food snack. Suppose that the sample standard deviation for the children who watched the sports celebrity-endorsed Walker Crisps commercial was 21.4 grams and the sample standard deviation for the children who watched the alternative food snack commercial was 12.8 grams.

- Assuming that the population variances are equal and $\alpha = 0.05$, is there evidence that the mean amount of Walker Crisps eaten was significantly higher for the children who watched the sports celebrity-endorsed Walker Crisps commercial?
- Assuming that the population variances are equal, construct a 95% confidence interval estimate of the difference between the mean amount of Walker Crisps eaten by children who watched the sports celebrity-endorsed Walker Crisps commercial and children who watched the alternative food snack commercial.
- Compare and discuss the results of (a) and (b).

10.9 A problem with a phone line that prevents a customer from receiving or making calls is upsetting to both the customer and the telecommunications company. The file **Phone** contains samples of 20 problems reported to two different offices of a telecommunications company and the time to clear these problems (in minutes) from the customers’ lines:

Central Office I Time to Clear Problems (minutes)

1.48	1.75	0.78	2.85	0.52	1.60	4.15	3.97	1.48	3.10
1.02	0.53	0.93	1.60	0.80	1.05	6.32	3.93	5.45	0.97

Central Office II Time to Clear Problems (minutes)

7.55	3.75	0.10	1.10	0.60	0.52	3.30	2.10	0.58	4.02
3.75	0.65	1.92	0.60	1.53	4.23	0.08	1.48	1.65	0.72

- Assuming that the population variances from both offices are equal, is there evidence of a difference in the mean waiting time between the two offices? (Use $\alpha = 0.05$.)
- Find the p -value in (a) and interpret its meaning.
- What other assumption is necessary in (a)?
- Assuming that the population variances from both offices are equal, construct and interpret a 95% confidence interval estimate of the difference between the population means in the two offices.

10.10 *Accounting Today* identified the top accounting firms in 10 geographic regions across the United States. Even though all 10 regions reported growth in 2012, the Southeast



and Gulf Coast regions reported the highest combined growths, with 7.94% and 9.92%, respectively. A characteristic description of the accounting firms in the Southeast and Gulf Coast regions included the number of partners in the firm. The file **AccountingPartners2** contains the number of partners. (Data extracted from bit.ly/11asPHm.)

- At the 0.05 level of significance, is there evidence of a difference between Southeast region accounting firms and Gulf Coast accounting firms with respect to the mean number of partners?
- Determine the p -value and interpret its meaning.
- What assumptions do you have to make about the two populations in order to justify the use of the t test?

10.11 An important feature of digital cameras is battery life—the number of shots that can be taken before the battery needs to be recharged. The file **Cameras** contains the battery life of 11 sub-compact cameras and 7 compact cameras. (Data extracted from “Cameras,” *Consumer Reports*, July 2012, pp. 42–44.)

- Assuming that the population variances from both types of digital cameras are equal, is there evidence of a difference in the mean battery life between the two types of digital cameras ($\alpha = 0.05$)?
- Determine the p -value in (a) and interpret its meaning.
- Assuming that the population variances from both types of digital cameras are equal, construct and interpret a 95% confidence interval estimate of the difference between the population mean battery life of the two types of digital cameras.

10.12 A bank with a branch located in a commercial district of a city has the business objective of developing an improved process for serving customers during the noon-to-1 P.M. lunch period. Management decides to first study the waiting time in the current process. The waiting time is defined as the number of minutes that elapses from when the customer enters the line until he or she reaches the teller window. Data are collected from a random sample of 15 customers and stored in **Bank1**. These data are:

4.21	5.55	3.02	5.13	4.77	2.34	3.54	3.20
4.50	6.10	0.38	5.12	6.46	6.19	3.79	

Suppose that another branch, located in a residential area, is also concerned with improving the process of serving customers in the noon-to-1 P.M. lunch period. Data are collected from a random sample of 15 customers and stored in **Bank2**. These data are:

9.66	5.90	8.02	5.79	8.73	3.82	8.01	8.35
10.49	6.68	5.64	4.08	6.17	9.91	5.47	

- Assuming that the population variances from both banks are equal, is there evidence of a difference in the mean waiting time between the two branches? (Use $\alpha = 0.05$.)
- Determine the p -value in (a) and interpret its meaning.
- In addition to equal variances, what other assumption is necessary in (a)?
- Construct and interpret a 95% confidence interval estimate of the difference between the population means in the two branches.

10.13 Repeat Problem 10.12 (a), assuming that the population variances in the two branches are not equal. Compare these results with those of Problem 10.12 (a).

10.14 As a member of the international strategic management team in your company, you are assigned the task of exploring

potential foreign market entry. As part of your initial investigation, you want to know if there is a difference between developed markets and emerging markets with respect to the time required to start a business. You select 15 developed countries and 15 emerging countries. The time required to start a business, defined as the number of days needed to complete the procedures to legally operate a business in these countries, is stored in **ForeignMarket**. (Data extracted from data.worldbank.org.)

- Assuming that the population variances for developed countries and emerging countries are equal, is there evidence of a difference in the mean time required to start a business between developed countries and emerging countries? (Use $\alpha = 0.05$.)
- Determine the p -value in (a) and interpret its meaning.
- In addition to equal variances, what other assumption is necessary in (a)?
- Construct a 95% confidence interval estimate of the difference between the population means of developed countries and emerging countries.

10.15 Repeat Problem 10.14 (a), assuming that the population variances from developed and emerging countries are not equal. Compare these results with those of Problem 10.14 (a).

10.16 An article appearing in *The Exponent*, an independent college newspaper published by the Purdue Student Publishing Foundation, reported that the average American college student spends 1 hour (60 minutes) on Facebook daily. (Data extracted from bit.ly/NQRCJQ.) But you wonder if there is a difference between males and females. You select a sample of 60 Facebook users (30 males and 30 females) at your college. The time spent on Facebook per day (in minutes) for these 60 users is stored in **FacebookTime2**.

- Assuming that the variances in the population of times spent on Facebook per day are equal, is there evidence of a difference in the mean time spent on Facebook per day between males and females? (Use a 0.05 level of significance.)
- In addition to equal variances, what other assumption is necessary in (a)?

10.17 Brand valuations are critical to CEOs, financial and marketing executives, security analysts, institutional investors, and others who depend on well-researched, reliable information needed for assessments, and comparisons in decision making. Millward Brown Optimor has developed the BrandZ Top 100 Most Valuable Global Brands for WPP, the world’s largest communications services group. Unlike other studies, the BrandZ Top 100 Most Valuable Global Brands fuses consumer measures of brand equity with financial measures to place a financial value on brands. The file **BrandZTechFin** contains the brand values for two sectors in the BrandZ Top 100 Most Valuable Global Brands for 2013: the technology sector and the financial institutions sector. (Data extracted from bit.ly/18OL5Mu.)

- Assuming that the population variances are equal, is there evidence of a difference between the technology sector and the financial institutions sector with respect to mean brand value? (Use $\alpha = .05$.)
- Repeat (a), assuming that the population variances are not equal.
- Compare the results of (a) and (b).

10.2 Comparing the Means of Two Related Populations

The hypothesis-testing procedures presented in Section 10.1 enable you to examine differences between the means of two *independent* populations. In this section, you will learn about a procedure for examining the mean difference between two populations when you collect sample data from populations that are related—that is, when results of the first population are *not* independent of the results of the second population.

There are two situations that involve related data: when you take repeated measurements from the same set of items or individuals or when you match items or individuals according to some characteristic. In either situation, you are interested in the *difference between the two related values* rather than the *individual values* themselves.

When you take **repeated measurements** on the same items or individuals, you assume that the same items or individuals will behave alike if treated alike. Your objective is to show that any differences between two measurements of the same items or individuals are due to different treatments that have been applied to the items or individuals. For example, when performing a taste-testing experiment comparing two beverages, you can use each person in the sample as his or her own control so that you can have *repeated measurements* on the same individual.

Another example of repeated measurements involves the pricing of the same goods from two different vendors. For example, have you ever wondered whether new textbook prices at a local college bookstore are different from the prices offered at a major online retailer? You could take two independent samples—that is, select two different sets of textbooks—and then use the hypothesis tests discussed in Section 10.1.

However, by random chance, the first sample may have many large-format hardcover textbooks and the second sample may have many small trade paperback books. This would imply that the first set of textbooks will always be more expensive than the second set of textbooks, regardless of where they are purchased. This observation means that using the Section 10.1 tests would not be a good choice. The better choice would be to use two related samples—that is, to determine the price of the *same* sample of textbooks at both the local bookstore and the online retailer.

The second situation that involves related data between populations is when you have **matched samples**. Here items or individuals are paired together according to some characteristic of interest. For example, in test marketing a product in two different advertising campaigns, a sample of test markets can be *matched* on the basis of the test market population size and/or demographic variables. By accounting for the differences in test market population size and/or demographic variables, you are better able to measure the effects of the two different advertising campaigns.

Regardless of whether you have matched samples or repeated measurements, the objective is to study the difference between two measurements by reducing the effect of the variability that is due to the items or individuals themselves. Table 10.3 shows the differences between the individual values for two related populations. To read this table, let $X_{11}, X_{12}, \dots, X_{1n}$ represent the n values from the first sample. And let $X_{21}, X_{22}, \dots, X_{2n}$ represent either the corresponding n matched values from a second sample or the corresponding n repeated measurements from the initial sample. Then D_1, D_2, \dots, D_n will represent the corresponding set of n difference scores such that

$$D_1 = X_{11} - X_{21}, D_2 = X_{12} - X_{22}, \dots, \text{and } D_n = X_{1n} - X_{2n}.$$

To test for the mean difference between two related populations, you treat the difference scores, each D_i , as values from a single sample.

TABLE 10.3

Determining the Difference Between Two Related Samples

 **Student Tip**
Which sample you define as group 1 will determine whether you will be doing a lower-tail test or an upper-tail test if you are conducting a one-tail test.

VALUE	SAMPLE		DIFFERENCE
	1	2	
1	X_{11}	X_{21}	$D_1 = X_{11} - X_{21}$
2	X_{12}	X_{22}	$D_2 = X_{12} - X_{22}$
\vdots	\vdots	\vdots	\vdots
i	X_{1i}	X_{2i}	$D_i = X_{1i} - X_{2i}$
\vdots	\vdots	\vdots	\vdots
n	X_{1n}	X_{2n}	$D_n = X_{1n} - X_{2n}$

Paired t Test

If you assume that the difference scores are randomly and independently selected from a population that is normally distributed, you can use the **paired t test for the mean difference** in related populations to determine whether there is a significant population mean difference. As with the one-sample t test developed in Section 9.2 [see Equation (9.2) on page 322], the paired t test statistic follows the t distribution with $n - 1$ degrees of freedom. Although the paired t test assumes that the population is normally distributed, since this test is robust, you can use this test as long as the sample size is not very small and the population is not highly skewed.

To test the null hypothesis that there is no difference in the means of two related populations:

$$H_0: \mu_D = 0 \text{ (where } \mu_D = \mu_1 - \mu_2\text{)}$$

against the alternative that the means are not the same:

$$H_1: \mu_D \neq 0$$

you compute the t_{STAT} test statistic using Equation (10.5).

PAIRED t TEST FOR THE MEAN DIFFERENCE

$$t_{STAT} = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}} \quad (10.5)$$

where

μ_D = hypothesized mean difference

$$\bar{D} = \frac{\sum_{i=1}^n D_i}{n}$$

$$S_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n - 1}}$$

The t_{STAT} test statistic follows a t distribution with $n - 1$ degrees of freedom.

For a two-tail test with a given level of significance, α , you reject the null hypothesis if the computed t_{STAT} test statistic is greater than the upper-tail critical value $t_{\alpha/2}$ from the t distribution, or, if the computed t_{STAT} test statistic is less than the lower-tail critical value $-t_{\alpha/2}$, from the t distribution. The decision rule is

Reject H_0 if $t_{STAT} > t_{\alpha/2}$

or if $t_{STAT} < -t_{\alpha/2}$;

otherwise, do not reject H_0 .

You can use the paired t test for the mean difference to investigate a question raised earlier in this section: Are new textbook prices at a local college bookstore different from the prices offered at a major online retailer?

In this repeated-measurements experiment, you use one set of textbooks. For each textbook, you determine the price at the local bookstore and the price at the online retailer. By determining the two prices for the same textbooks, you can reduce the variability in the prices compared with what would occur if you used two independent sets of textbooks. This approach focuses on the differences between the prices of the same textbooks offered by the two retailers.

You collect data by conducting an experiment from a sample of $n = 16$ textbooks used primarily in business school courses during a recent semester at a local college. You determine the college bookstore price and the online price (which includes shipping costs, if any). You organize and store the data in **BookPrices**. Table 10.4 shows the results. Notice that each row of the table shows the bookstore price and online retailer price for a specific book.

TABLE 10.4

Prices of Textbooks at the College Bookstore and at the Online Retailer

Author	Title	Bookstore	Online
Bade	<i>Foundations of Microeconomics 6/e</i>	200.00	121.49
Brigham	<i>Financial Management 13/e</i>	304.00	235.88
Clauretie	<i>Real Estate Finance: Theory and Practice</i>	179.35	107.61
Foner	<i>Give Me Liberty! (Brief) Vol. 2 3/e</i>	72.00	59.99
Garrison	<i>Managerial Accounting</i>	277.15	146.99
Grewal	<i>M: Marketing 3/e</i>	73.75	63.49
Hill	<i>Global Business Today</i>	171.65	138.99
Lafore	<i>Object-Oriented Programming in C++</i>	65.00	42.26
Lank	<i>Modern Real Estate Practice 11/e</i>	47.45	65.99
Meyer	<i>Entrepreneurship</i>	106.00	37.83
Mitchell	<i>Public Affairs in the Nation and New York</i>	55.95	102.99
Pindyck	<i>Microeconomics 8/e</i>	224.40	144.99
Robbins	<i>Organizational Behavior 15/e</i>	223.20	179.39
Ross	<i>Fundamentals of Corporate Finance 9/e</i>	250.65	191.49
Schneier	<i>New York Politics: Tale of Two States</i>	34.95	28.66
Wilson	<i>American Government: The Essentials 12/e</i>	172.65	108.49

Your objective is to determine whether there is any difference between the mean textbook price at the college bookstore and at the online retailer. In other words, is there evidence that the mean price is different between the two textbook sellers? Thus, the null and alternative hypotheses are

$H_0: \mu_D = 0$ (There is no difference in the mean price between the college bookstore and the online retailer.)

$H_1: \mu_D \neq 0$ (There is a difference in the mean price between the college bookstore and the online retailer.)

Choosing the level of significance $\alpha = 0.05$ and assuming that the differences are normally distributed, you use the paired t test [Equation (10.5)]. For a sample of $n = 16$ textbooks, there are $n - 1 = 15$ degrees of freedom. Using Table E.3, the decision rule is

Reject H_0 if $t_{STAT} > 2.1314$
or if $t_{STAT} < -2.1314$;
otherwise, do not reject H_0 .

For the $n = 16$ differences (see Table 10.4), the sample mean difference is

$$\bar{D} = \frac{\sum_{i=1}^n D_i}{n} = \frac{681.62}{16} = 42.6013$$

and

$$S_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1}} = 43.797$$

From Equation (10.5) on page 360,

$$t_{STAT} = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}} = \frac{42.6013 - 0}{\frac{43.797}{\sqrt{16}}} = 3.8908$$

Because $t_{STAT} = 3.8908 > 2.1314$, you reject the null hypothesis, H_0 (see Figure 10.8). There is evidence of a difference in the mean price of textbooks purchased at the college bookstore and the online retailer. You can conclude that the mean price is higher at the college bookstore than at the online retailer.

FIGURE 10.8

Two-tail paired t test at the 0.05 level of significance with 15 degrees of freedom

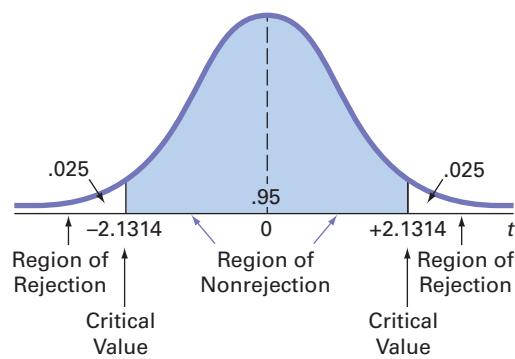


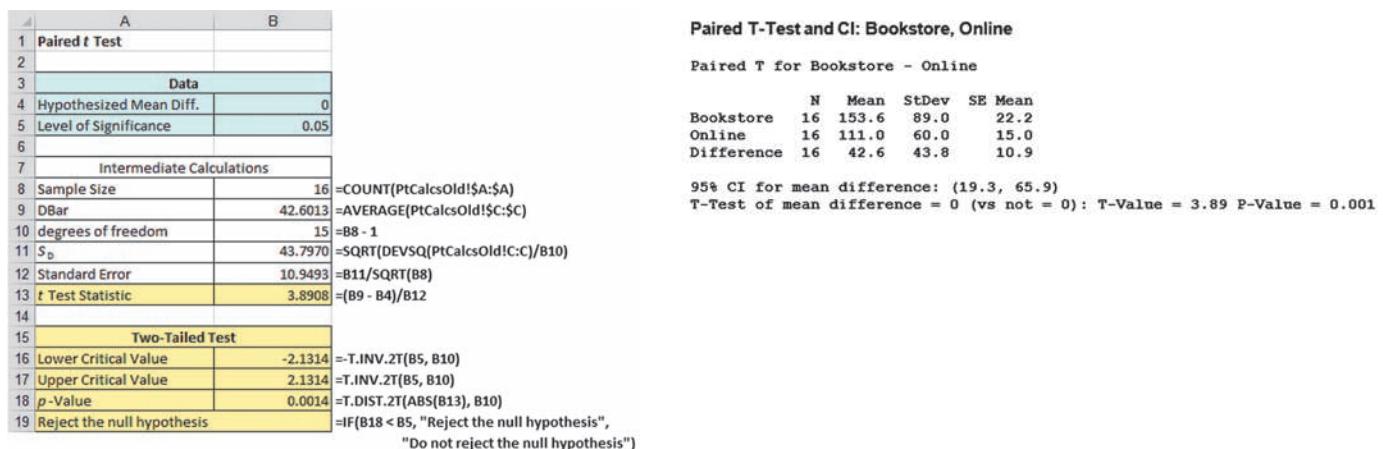
Figure 10.9 presents the results for this example, computing both the t test statistic and the p -value. Because the p -value $= 0.0014 < \alpha = 0.05$, you reject H_0 . The p -value indicates that if the two sources for textbooks have the same population mean price, the probability that one source would have a sample mean \$42.60 more than the other is 0.0014. Because this probability is less than $\alpha = 0.05$, you conclude that there is evidence to reject the null hypothesis.

To evaluate the validity of the assumption of normality, you construct a boxplot of the differences, as shown in Figure 10.10.

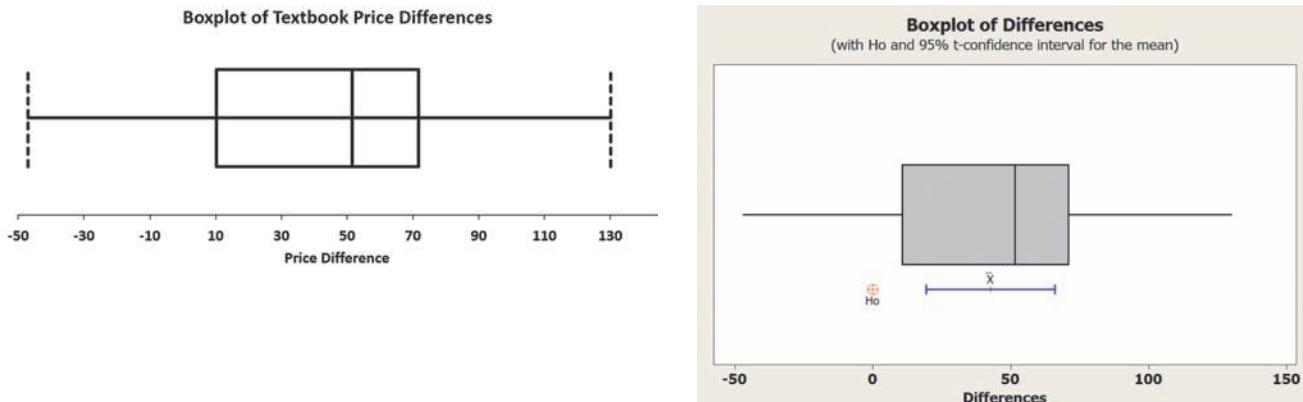
The Figure 10.10 boxplots show approximate symmetry and look similar to the boxplot for the normal distribution displayed in Figure 3.5 on page 125. Thus, the distribution of textbook price differences does not greatly contradict the underlying assumption of normality. If a boxplot, histogram, or normal probability plot reveals that the assumption of underlying normality in the population is severely violated, then the t test may be inappropriate, especially if

FIGURE 10.9

Excel and Minitab paired *t* test results for the textbook price data

**FIGURE 10.10**

Excel and Minitab boxplots for the textbook price differences



the sample size is small. If you believe that the *t* test is inappropriate, you can use either a *non-parametric* procedure that does not make the assumption of underlying normality (see online Section 12.8 or references 1 and 2) or make a data transformation (see reference 6) and then recheck the assumptions to determine whether you should use the *t* test.

EXAMPLE 10.2

Paired *t* Test of Pizza Delivery Times

Recall from Example 10.1 on page 351 that a local pizza restaurant situated across the street from your college campus advertises that it delivers to the dormitories faster than the local branch of a national pizza chain. In order to determine whether this advertisement is valid, you and some friends decided to order 10 pizzas from the local pizza restaurant and 10 pizzas from the national chain. In fact, each time you ordered a pizza from the local pizza restaurant, at the same time, your friends ordered a pizza from the national pizza chain. Thus, you have matched samples. For each of the 10 times that pizzas were ordered, you have one measurement from the local pizza restaurant and one from the national chain. At the 0.05 level of significance, is the mean delivery time for the local pizza restaurant less than the mean delivery time for the national pizza chain?

(continued)

SOLUTION Use the paired *t* test to analyze the Table 10.5 data (stored in **PizzaTime**). Figure 10.11 shows the paired *t* test results for the pizza delivery data.

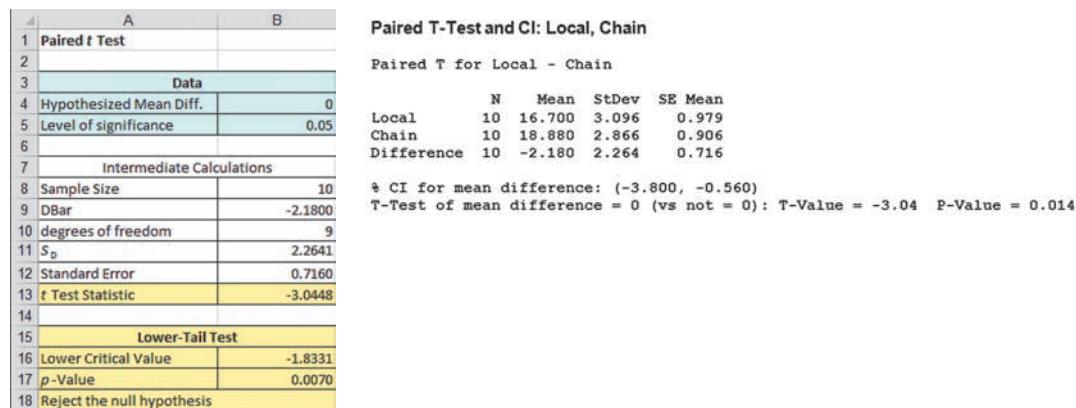
TABLE 10.5

Delivery Times
for Local Pizza
Restaurant and
National Pizza Chain

Time	Local	Chain	Difference
1	16.8	22.0	-5.2
2	11.7	15.2	-3.5
3	15.6	18.7	-3.1
4	16.7	15.6	1.1
5	17.5	20.8	-3.3
6	18.1	19.5	-1.4
7	14.1	17.0	-2.9
8	21.8	19.5	2.3
9	13.9	16.5	-2.6
10	20.8	24.0	-3.2
			-21.8

FIGURE 10.11

Excel and Minitab
paired *t* test results for
the pizza delivery data



The null and alternative hypotheses are

$H_0: \mu_D \geq 0$ (Mean difference in the delivery time between the local pizza restaurant and the national pizza chain is greater than or equal to 0.)

$H_1: \mu_D < 0$ (Mean difference in the delivery time between the local pizza restaurant and the national pizza chain is less than 0.)

Choosing the level of significance $\alpha = 0.05$ and assuming that the differences are normally distributed, you use the paired *t* test [Equation (10.5) on page 360]. For a sample of $n = 10$ delivery times, there are $n - 1 = 9$ degrees of freedom. Using Table E.3, the decision rule is

Reject H_0 if $t_{STAT} < -t_{0.05} = -1.8331$;
otherwise, do not reject H_0 .

To illustrate the computations, for $n = 10$ differences (see Table 10.5), the sample mean difference is

$$\bar{D} = \frac{\sum_{i=1}^n D_i}{n} = \frac{-21.8}{10} = -2.18$$

and the sample standard deviation of the difference is

$$S_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n - 1}} = 2.2641$$

From Equation (10.5) on page 360,

$$t_{STAT} = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}} = \frac{-2.18 - 0}{\frac{2.2641}{\sqrt{10}}} = -3.0448$$

Because $t_{STAT} = -3.0448$ is less than -1.8331 , you reject the null hypothesis, H_0 (the p -value is $0.0070 < 0.05$). There is evidence that the mean delivery time is lower for the local pizza restaurant than for the national pizza chain.

This conclusion differs from the conclusion you reached on page 353 for Example 10.1 when you used the pooled-variance t test for these data. By pairing the delivery times, you are able to focus on the differences between the two pizza delivery services and not the variability created by ordering pizzas at different times of day. The paired t test is a more powerful statistical procedure that reduces the variability in the delivery time because you are controlling for the time of day the pizza was ordered.

Confidence Interval Estimate for the Mean Difference

Instead of or in addition to testing for the mean difference between two related populations, you can use Equation (10.6) to construct a confidence interval estimate for the population mean difference.

CONFIDENCE INTERVAL ESTIMATE FOR THE MEAN DIFFERENCE

$$\bar{D} \pm t_{\alpha/2} \frac{S_D}{\sqrt{n}}$$

or

$$\bar{D} - t_{\alpha/2} \frac{S_D}{\sqrt{n}} \leq \mu_D \leq \bar{D} + t_{\alpha/2} \frac{S_D}{\sqrt{n}} \quad (10.6)$$

where $t_{\alpha/2}$ is the critical value of the t distribution, with $n - 1$ degrees of freedom, for an area of $\alpha/2$ in the upper tail.

Recall the example comparing textbook prices on page 361. Using Equation (10.6), $\bar{D} = 42.6013$, $S_D = 43.797$, $n = 16$, and $t_{\alpha/2} = 2.1314$ (for 95% confidence and $n - 1 = 15$ degrees of freedom),

$$42.6013 \pm (2.1314) \frac{43.797}{\sqrt{16}}$$

$$42.6013 \pm 23.3373$$

$$19.264 \leq \mu_D \leq 65.9386$$

Thus, with 95% confidence, you estimate that the population mean difference in textbook prices between the college bookstore and the online retailer is between \$19.26 and \$65.94.

Because the interval estimate does not contain zero, using the 0.05 level of significance and a two-tail test, you can conclude that there is evidence of a difference in the mean prices of textbooks at the college bookstore and the online retailer. Since both the lower and upper limits of the confidence interval are above 0, you can conclude that the mean price is higher at the college bookstore than the online retailer.

Problems for Section 10.2

LEARNING THE BASICS

10.18 An experimental design for a paired t test has 20 pairs of identical twins. How many degrees of freedom are there in this t test?

10.19 Fifteen volunteers are recruited to participate in an experiment. A measurement is made (such as blood pressure) before each volunteer is asked to read a particularly upsetting passage from a book and after each volunteer reads the passage from the book. In the analysis of the data collected from this experiment, how many degrees of freedom are there in the test?

APPLYING THE CONCEPTS

 **10.20** Nine experts rated two brands of Colombian coffee in a taste-testing experiment. A rating on a 7-point scale (1 = extremely unpleasing, 7 = extremely pleasing) is given for each of four characteristics: taste, aroma, richness, and acidity. The following data stored in **Coffee** contain the ratings accumulated over all four characteristics:

EXPERT	BRAND	
	A	B
C.C.	24	26
S.E.	27	27
E.G.	19	22
B.L.	24	27
C.M.	22	25
C.N.	26	27
G.N.	27	26
R.M.	25	27
P.V.	22	23

- a. At the 0.05 level of significance, is there evidence of a difference in the mean ratings between the two brands?
- b. What assumption is necessary about the population distribution in order to perform this test?
- c. Determine the p -value in (a) and interpret its meaning.
- d. Construct and interpret a 95% confidence interval estimate of the difference in the mean ratings between the two brands.

10.21 How do the ratings of TV and phone services compare?

The file **Telecom** contains the rating of 13 different providers. (Data extracted from “Ratings: TV, Phone, and Internet Services,” *Consumer Reports*, May 2012, p. 25.)

- a. At the 0.05 level of significance, is there evidence of a difference in the mean service rating between TV and phone services?
- b. What assumption is necessary about the population distribution in order to perform this test?
- c. Use a graphical method to evaluate the validity of the assumption in (a).
- d. Construct and interpret a 95% confidence interval estimate of the difference in the mean service rating between TV and phone services.

10.22 Super Target versus Walmart: Who has the lowest prices?

Given Walmart’s slogan “Save Money—Live Better,” you suspect that Walmart does. The prices of 33 foods were compared (data extracted from “Supermarket Showdown,” *The Palm Beach Post*, February 13, 2011, pp. 1F, 2F) and the results are stored in **TargetWalmart**.

- a. At the 0.05 level of significance, is there evidence that the mean price of items is higher at Super Target than at Walmart?
- b. What assumption is necessary about the population distribution in order to perform this test?
- c. Find the p -value in (a) and interpret its meaning.

10.23 What motivates employees? The Great Place to Work Institute evaluated nonfinancial factors both globally and in the United States. (Data extracted from L. Petrecca, “Tech Companies Top List of ‘Great Workplaces,’” *USA Today*, October 31, 2011, p. 7B.) The results, which indicate the importance rating of each factor, are stored in **Motivation**.

- a. At the 0.05 level of significance, is there evidence of a difference in the mean rating between global and U.S. employees?
- b. What assumption is necessary about the population distribution in order to perform this test?
- c. Use a graphical method to evaluate the validity of the assumption in (b).

10.24 Multiple myeloma, or blood plasma cancer, is characterized by increased blood vessel formulation (angiogenesis) in the bone marrow that is a predictive factor in survival. One treatment approach used for multiple myeloma is stem cell transplantation with the patient’s own stem cells. The data stored in **Myeloma**, and shown on page 367 represent the bone marrow microvessel density for patients who had a complete response to the stem cell transplant (as measured by blood and urine tests). The measurements were taken immediately prior to the stem cell transplant and at the time the complete response was determined.

Patient	Before	After
1	158	284
2	189	214
3	202	101
4	353	227
5	416	290
6	426	176
7	441	290

Data extracted from S. V. Rajkumar, R. Fonseca, T. E. Witzig, M. A. Gertz, and P. R. Greipp, "Bone Marrow Angiogenesis in Patients Achieving Complete Response After Stem Cell Transplantation for Multiple Myeloma," *Leukemia* 13 (1999): 469–472.

- a. At the 0.05 level of significance, is there evidence that the mean bone marrow microvessel density is higher before the stem cell transplant than after the stem cell transplant?
- b. Interpret the meaning of the p -value in (a).
- c. Construct and interpret a 95% confidence interval estimate of the mean difference in bone marrow microvessel density before and after the stem cell transplant.

- d. What assumption is necessary about the population distribution in order to perform the test in (a)?

10.25 To assess the effectiveness of a cola video ad, a random sample of 38 individuals from a target audience was selected to participate in a copy test. Participants viewed two ads, one of which was the ad being tested. Participants then answered a series of questions about how much they liked the ads. An adindex measure was created and stored in **Adindex**; the higher the adindex value, the more likeable the ad. Compute descriptive statistics and perform a paired t test. State your findings and conclusions in a report. (use the 0.05 level of significance.)

10.26 The file **Concrete1** contains the compressive strength, in thousands of pounds per square inch (psi), of 40 samples of concrete taken two and seven days after pouring. (Data extracted from O. Carrillo-Gamboa and R. F. Gunst, "Measurement-Error-Model Collinearities," *Technometrics*, 34 (1992): 454–464.)

- a. At the 0.01 level of significance, is there evidence that the mean strength is lower at two days than at seven days?
- b. What assumption is necessary about the population distribution in order to perform this test?
- c. Find the p -value in (a) and interpret its meaning.

10.3 Comparing the Proportions of Two Independent Populations

Often, you need to make comparisons and analyze differences between two population proportions. You can perform a test for the difference between two proportions selected from independent populations by using two different methods. This section presents a procedure whose test statistic, Z_{STAT} , is approximated by a standardized normal distribution. In Section 12.1, a procedure whose test statistic, χ^2_{STAT} , is approximated by a chi-square distribution is used. As explained in the latter section, the results from these two tests are equivalent.

Z Test for the Difference Between Two Proportions

In evaluating differences between two population proportions, you can use a **Z test for the difference between two proportions**. The Z_{STAT} test statistic is based on the difference between two sample proportions ($p_1 - p_2$). This test statistic, given in Equation (10.7), approximately follows a standardized normal distribution for large enough sample sizes.

Z TEST FOR THE DIFFERENCE BETWEEN TWO PROPORTIONS

$$Z_{STAT} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (10.7)$$

with

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} \quad p_1 = \frac{X_1}{n_1} \quad p_2 = \frac{X_2}{n_2}$$

(continued)

 **Student Tip**

Do not confuse this use of the Greek letter pi, π , to represent the population proportion with the mathematical constant that is the ratio of the circumference to a diameter of a circle—approximately 3.14159—which is also known by the same Greek letter.

where

- p_1 = proportion of items of interest in sample 1
- X_1 = number of items of interest in sample 1
- n_1 = sample size of sample 1
- π_1 = proportion of items of interest in population 1
- p_2 = proportion of items of interest in sample 2
- X_2 = number of items of interest in sample 2
- n_2 = sample size of sample 2
- π_2 = proportion of items of interest in population 2
- \bar{p} = pooled estimate of the population proportion of items of interest

The Z_{STAT} test statistic approximately follows a standardized normal distribution.

The null hypothesis in the Z test for the difference between two proportions states that the two population proportions are equal ($\pi_1 = \pi_2$). Because the pooled estimate for the population proportion is based on the null hypothesis, you combine, or pool, the two sample proportions to compute \bar{p} , an overall estimate of the common population proportion. This estimate is equal to the number of items of interest in the two samples ($X_1 + X_2$) divided by the total sample size from the two samples ($n_1 + n_2$).

As shown in the following table, you can use this Z test for the difference between population proportions to determine whether there is a difference in the proportion of items of interest in the two populations (two-tail test) or whether one population has a higher proportion of items of interest than the other population (one-tail test):

Two-Tail Test	One-Tail Test	One-Tail Test
$H_0: \pi_1 = \pi_2$	$H_0: \pi_1 \geq \pi_2$	$H_0: \pi_1 \leq \pi_2$
$H_1: \pi_1 \neq \pi_2$	$H_1: \pi_1 < \pi_2$	$H_1: \pi_1 > \pi_2$

where

π_1 = proportion of items of interest in population 1

π_2 = proportion of items of interest in population 2

To test the null hypothesis that there is no difference between the proportions of two independent populations:

$$H_0: \pi_1 = \pi_2$$

against the alternative that the two population proportions are not the same:

$$H_1: \pi_1 \neq \pi_2$$

you use the Z_{STAT} test statistic, given by Equation (10.7). For a given level of significance, α , you reject the null hypothesis if the computed Z_{STAT} test statistic is greater than the upper-tail critical value from the standardized normal distribution or if the computed Z_{STAT} test statistic is less than the lower-tail critical value from the standardized normal distribution.

To illustrate the use of the Z test for the equality of two proportions, suppose that you are the manager of T.C. Resort Properties, a collection of five upscale resort hotels located on two tropical islands. On one of the islands, T.C. Resort Properties has two hotels, the Beachcomber and the Windsurfer. Using the DCOVA problem solving approach, you have defined the business objective as improving the return rate of guests at the Beachcomber and the Windsurfer hotels. On the survey completed by hotel guests upon or after their departure, one question asked is whether the guest is likely to return to the hotel. Responses to this and other questions were collected from 227 guests at the Beachcomber and 262 guests at the Windsurfer. The results for this

question indicated that 163 of 227 guests at the Beachcomber responded yes, they were likely to return to the hotel and 154 of 262 guests at the Windsurfer responded yes, they were likely to return to the hotel. At the 0.05 level of significance, is there evidence of a significant difference in guest satisfaction (as measured by the likelihood to return to the hotel) between the two hotels?

The null and alternative hypotheses are

$$H_0: \pi_1 = \pi_2 \text{ or } \pi_1 - \pi_2 = 0$$

$$H_1: \pi_1 \neq \pi_2 \text{ or } \pi_1 - \pi_2 \neq 0$$

Using the 0.05 level of significance, the critical values are -1.96 and $+1.96$ (see Figure 10.12), and the decision rule is

Reject H_0 if $Z_{STAT} < -1.96$

or if $Z_{STAT} > +1.96$;

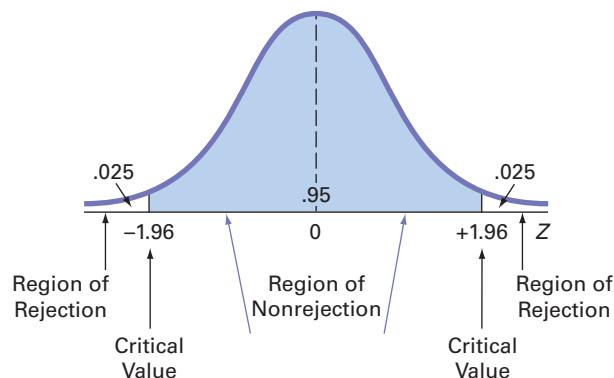
otherwise, do not reject H_0 .

Using Equation (10.7) on page 367,

$$Z_{STAT} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

FIGURE 10.12

Regions of rejection and nonrejection when testing a hypothesis for the difference between two proportions at the 0.05 level of significance



where

$$p_1 = \frac{X_1}{n_1} = \frac{163}{227} = 0.7181 \quad p_2 = \frac{X_2}{n_2} = \frac{154}{262} = 0.5878$$

and

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{163 + 154}{227 + 262} = \frac{317}{489} = 0.6483$$

so that

$$\begin{aligned} Z_{STAT} &= \frac{(0.7181 - 0.5878) - (0)}{\sqrt{0.6483(1 - 0.6483)\left(\frac{1}{227} + \frac{1}{262}\right)}} \\ &= \frac{0.1303}{\sqrt{(0.228)(0.0082)}} \\ &= \frac{0.1303}{\sqrt{0.00187}} \\ &= \frac{0.1303}{0.0432} = +3.0088 \end{aligned}$$

Using the 0.05 level of significance, you reject the null hypothesis because $Z_{STAT} = +3.0088 > +1.96$. The p -value is 0.0026 (computed using Table E.2 or from Figure 10.13) and indicates that if the null hypothesis is true, the probability that a Z_{STAT} test statistic is less than -3.0088 is 0.0013, and, similarly, the probability that a Z_{STAT} test statistic is greater than $+3.0088$ is 0.0013. Thus, for this two-tail test, the p -value is $0.0013 + 0.0013 = 0.0026$. Because $0.0026 < \alpha = 0.05$, you reject the null hypothesis. There is evidence to conclude that the two hotels are significantly different with respect to guest satisfaction; a greater proportion of guests are willing to return to the Beachcomber than to the Windsurfer.

FIGURE 10.13

Excel and Minitab Z test results for the difference between two proportions for the hotel guest satisfaction problem

A	B
1 Z Test for Differences in Two Proportions	
2	
3 Data	
4 Hypothesized Difference	0
5 Level of Significance	0.05
6 Group 1	
7 Number of Successes	163
8 Sample Size	227
9 Group 2	
10 Number of Successes	154
11 Sample Size	262
12	
13 Intermediate Calculations	
14 Group 1 Proportion	0.7181 =B7/B8
15 Group 2 Proportion	0.5878 =B10/B11
16 Difference in Two Proportions	0.1303 =B14 - B15
17 Average Proportion	0.6483 =(B7 + B10)/(B8 + B11)
18 Z Test Statistic	3.0088 =(B16 - B4)/SQRT(B17 * (1 - B17) * (1/B8 + 1/B11))
19	
20 Two-Tail Test	
21 Lower Critical Value	-1.9600 =NORM.S.INV(B5/2)
22 Upper Critical Value	1.9600 =NORM.S.INV(1 - B5/2)
23 p-Value	0.0026 =2 * (1 - NORM.S.DIST(ABS(B18), TRUE))
24 Reject the null hypothesis	=IF(B23 < B5, "Reject the null hypothesis", "Do not reject the null hypothesis")

Test and CI for Two Proportions

Sample	X	N	Sample p
1	163	227	0.718062
2	154	262	0.587786

Difference = p (1) - p (2)
Estimate for difference: 0.130275
95% CI for difference: (0.0467379, 0.213813)
Test for difference = 0 (vs not = 0): Z = 3.01 P-Value = 0.003

Fisher's exact test: P-Value = 0.003

EXAMPLE 10.3

Testing for the Difference Between Two Proportions

Are men less likely than women to shop for bargains? A survey reported that when going shopping, 24% of men (181 of 756 sampled) and 34% of women (275 of 809 sampled) go for bargains. (Data extracted from “Brands More Critical for Dads,” *USA Today*, July 21, 2011, p. 1C.) At the 0.05 level of significance, is the proportion of men who shop for bargains less than the proportion of women who shop for bargains?

SOLUTION Because you want to know whether there is evidence that the proportion of men who shop for bargains is *less* than the proportion of women who shop for bargains, you have a one-tail test. The null and alternative hypotheses are

$H_0: \pi_1 \geq \pi_2$ (The proportion of men who shop for bargains is greater than or equal to the proportion of women who shop for bargains.)

$H_1: \pi_1 < \pi_2$ (The proportion of men who shop for bargains is less than the proportion of women who shop for bargains.)

Using the 0.05 level of significance, for the one-tail test in the lower tail, the critical value is +1.645. The decision rule is

Reject H_0 if $Z_{STAT} < -1.645$;
otherwise, do not reject H_0 .

Using Equation (10.7) on page 367,

$$Z_{STAT} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$p_1 = \frac{X_1}{n_1} = \frac{181}{756} = 0.2394 \quad p_2 = \frac{X_2}{n_2} = \frac{275}{809} = 0.3399$$

and

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{181 + 275}{756 + 809} = \frac{456}{1565} = 0.2914$$

so that

$$\begin{aligned} Z_{STAT} &= \frac{(0.2394 - 0.3399) - (0)}{\sqrt{0.2914(1 - 0.2914)\left(\frac{1}{756} + \frac{1}{809}\right)}} \\ &= \frac{-0.1005}{\sqrt{(0.2065)(0.00256)}} \\ &= \frac{-0.1005}{\sqrt{0.00053}} \\ &= \frac{-0.1005}{0.0230} = -4.37 \end{aligned}$$

Using the 0.05 level of significance, you reject the null hypothesis because $Z_{STAT} = -4.37 < -1.645$. The p -value is 0.0000. Therefore, if the null hypothesis is true, the probability that a Z_{STAT} test statistic is less than -4.37 is 0.0000 (which is less than $\alpha = 0.05$). You conclude that there is evidence that the proportion of men who shop for bargains is less than the proportion of women who shop for bargains.

Confidence Interval Estimate for the Difference Between Two Proportions

Instead of or in addition to testing for the difference between the proportions of two independent populations, you can construct a confidence interval estimate for the difference between the two proportions using Equation (10.8).

CONFIDENCE INTERVAL ESTIMATE FOR THE DIFFERENCE BETWEEN TWO PROPORTIONS

$$(p_1 - p_2) \pm Z_{\alpha/2} \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

or

$$\begin{aligned} (p_1 - p_2) - Z_{\alpha/2} \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} &\leq (\pi_1 - \pi_2) \\ (p_1 - p_2) + Z_{\alpha/2} \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} &\quad (10.8) \end{aligned}$$

To construct a 95% confidence interval estimate for the population difference between the proportion of guests who would return to the Beachcomber and who would return to the Windsurfer, you use the results on page 369 or from Figure 10.13 on page 370:

$$p_1 = \frac{X_1}{n_1} = \frac{163}{227} = 0.7181 \quad p_2 = \frac{X_2}{n_2} = \frac{154}{262} = 0.5878$$

Using Equation (10.8),

$$(0.7181 - 0.5878) \pm (1.96) \sqrt{\frac{0.7181(1 - 0.7181)}{227} + \frac{0.5878(1 - 0.5878)}{262}}$$

$$0.1303 \pm (1.96)(0.0426)$$

$$0.1303 \pm 0.0835$$

$$0.0468 \leq (\pi_1 - \pi_2) \leq 0.2138$$

Thus, you have 95% confidence that the difference between the population proportion of guests who would return to the Beachcomber and the Windsurfer is between 0.0468 and 0.2138. In percentages, the difference is between 4.68% and 21.38%. Guest satisfaction is higher at the Beachcomber than at the Windsurfer.

Problems for Section 10.3

LEARNING THE BASICS

10.27 Let $n_1 = 100$, $X_1 = 50$, $n_2 = 100$, and $X_2 = 30$.

- a. At the 0.05 level of significance, is there evidence of a significant difference between the two population proportions?
- b. Construct a 95% confidence interval estimate for the difference between the two population proportions.

10.28 Let $n_1 = 100$, $X_1 = 45$, $n_2 = 50$, and $X_2 = 25$.

- a. At the 0.01 level of significance, is there evidence of a significant difference between the two population proportions?
- b. Construct a 99% confidence interval estimate for the difference between the two population proportions.

APPLYING THE CONCEPTS

10.29 A survey of 1,085 adults asked, “Do you enjoy shopping for clothing for yourself?” The results indicated that 51% of the females enjoyed shopping for clothing for themselves as compared to 44% of the males. (Data extracted from “Split Decision on Clothes Shopping,” *USA Today*, January 28, 2011, p. 1B.) The sample sizes of males and females were not provided. Suppose that of 542 males, 238 said that they enjoyed shopping for clothing for themselves while of 543 females, 276 said that they enjoyed shopping for clothing for themselves.

- a. Is there evidence of a difference between males and females in the proportion who enjoy shopping for clothing for themselves at the 0.01 level of significance?
- b. Find the p -value in (a) and interpret its meaning.
- c. Construct and interpret a 99% confidence interval estimate for the difference between the proportion of males and females who enjoy shopping for clothing for themselves.
- d. What are your answers to (a) through (c) if 218 males enjoyed shopping for clothing for themselves?

10.30 Do social recommendations increase ad effectiveness? A study of online video viewers compared viewers who arrived at an advertising video for a particular brand by following a social media recommendation link to viewers who arrived at the same video by web browsing. Data were collected on whether the viewer could correctly recall the brand being advertised after seeing the video. The results were:

ARRIVAL METHOD	CORRECTLY RECALLED THE BRAND	
	Yes	No
Recommendation	407	150
Browsing	193	91

Source: Data extracted from “Social Ad Effectiveness: An Unruly White Paper,” www.unrulymedia.com, January 2012, p.3.

- a. Set up the null and alternative hypotheses to try to determine whether brand recall is higher following a social media recommendation than with only web browsing.
- b. Conduct the hypothesis test defined in (a), using the 0.05 level of significance.
- c. Does the result of your test in (b) make it appropriate to claim that brand recall is higher following a social media recommendation than by web browsing?

10.31 A/B testing is a testing method that businesses use to test different designs and formats of a web page to determine whether a new web page is more effective than a current web page. Web designers at TravelTips.com tested a new call to action button on its web page. Every visitor to the web page was randomly shown

either the original call to action button (the control) or the new call to action button. The metric used to measure success was the download rate: the number of people who downloaded the file divided by the number of people who saw that particular call to action button. The experiment yielded the following results:

Variations	Downloads	Visitors
Original call to action button	351	3,642
New call to action button	485	3,556

- a. What is the proportion (download rate) of visitors who saw the original call to action button and downloaded the file?
- b. What is the proportion (download rate) of visitors who saw the new call to action button and downloaded the file?
- c. At the 0.05 level of significance, is there evidence that the new call to action button is more effective than the original?

 **10.32** The consumer research firm Scarborough analyzed the 10% of American adults that are either “Superbanked” or “Unbanked.” Superbanked consumers are defined as U.S. adults who live in a household that has multiple asset accounts at financial institutions, as well as some additional investments; Unbanked consumers are U.S. adults who live in a household that does not use a bank or credit union. By finding the 5% of Americans that are Superbanked, Scarborough identifies financially savvy consumers who might be open to diversifying their financial portfolios; by identifying the Unbanked, Scarborough provides insight into the ultimate prospective client for banks and financial institutions. As part of its analysis, Scarborough reported that 93% of Superbanked consumers use credit cards in the past three months as compared to 23% of Unbanked consumers. (Data extracted from bit.ly/QIABwO.) Suppose that these results were based on 1,000 Superbanked consumers and 1,000 Unbanked consumers.

- a. At the 0.01 level of significance, is there evidence of a significant difference between the Superbanked and the Unbanked with respect to the proportion that use credit cards?
- b. Find the p -value in (a) and interpret its meaning.
- c. Construct and interpret a 99% confidence interval estimate for the difference between the Superbanked and the Unbanked with respect to the proportion that use credit cards.

10.33 What social media tools do marketers commonly use? A survey by Social Media Examiner (data extracted from “2012 Social Media Marketing Industry Report,” April 2012, p. 27) of B2B marketers (marketers that focus primarily on attracting businesses) and B2C marketers (marketers that primarily target consumers) reported that 87% of B2B marketers and 59% of B2C marketers

commonly use LinkedIn as a social media tool. The study also revealed that 56% of B2B marketers and 59% of B2C marketers commonly use YouTube or other video as a social media tool. Suppose the survey was based on 500 B2B marketers and 500 B2C marketers.

- a. At the 0.05 level of significance, is there evidence of a difference between B2B marketers and B2C marketers in the proportion that commonly use LinkedIn as a social media tool?
- b. Find the p -value in (a) and interpret its value.
- c. At the 0.05 level of significance, is there evidence of a difference between B2B marketers and B2C marketers in the proportion that commonly use YouTube or other video as a social media tool?

10.34 Does gamification motivate customer research management (CRM) utilization? Gamification is the use of game mechanics to motivate, modify, or reward distinct behaviors. In the context of sales effectiveness, it is deployed to encourage both sales accomplishments and non-sales activities. A survey of end-user sales organizations indicates that 31 of 37 gamification-user organizations provide mobile access to CRM, whereas 138 of 275 non-gamification-user organizations provide mobile access to CRM. (Data extracted from v1.aberdeen.com/launch/report/research_briefs/8346-RB-gamification-sales-effectiveness.asp?lan=US.)

- a. At the 0.05 level of significance, is there evidence of a difference between gamification-user sales organizations and non-gamification-user sales organizations in the proportion that provide mobile access to CRM?
- b. Find the p -value in (a) and interpret its meaning.

10.35 One of the most impressive, innovative advances in online fundraising over the past decade is the rise of crowd-funding websites. While features differ from site to site, crowd-funding sites are websites that allow you to set up an online fundraising campaign based around a fundraising page, and accept money directly from that page using the website’s own credit card processor. Kickstarter, one crowd-funding website, reported that 316 of 831 *technology* crowd-funding projects were successfully launched in the past year and 923 of 2,796 *games* crowd-funding projects were successfully launched in the past year. (Data extracted from kickstarter.com/hello?ref=nav.)

- a. Is there evidence of a significant difference in the proportion of *technology* crowd-funding projects and *games* crowd-funding projects that were successful? (Use $\alpha = 0.05$.)
- b. Determine the p -value in (a) and interpret its meaning.
- c. Construct and interpret a 95% confidence interval estimate for the difference between the proportion of *technology* crowd-funding projects and *games* crowd-funding projects that are successful.

10.4 F Test for the Ratio of Two Variances

Often you need to determine whether two independent populations have the same variability. By testing variances, you can detect differences in the variability in two independent populations. One important reason to test for the difference between the variances of two populations is to determine whether to use the pooled-variance t test (which assumes equal variances) or the separate-variance t test (which does not assume equal variances) when comparing the means of two independent populations.

The test for the difference between the variances of two independent populations is based on the ratio of the two sample variances. If you assume that each population is normally distributed, then the sampling distribution of the ratio S_1^2/S_2^2 is distributed as an *F* distribution (see Table E.5). The critical values of the ***F* distribution** in Table E.5 depend on the degrees of freedom in the two samples. The degrees of freedom in the numerator of the ratio are for the first sample, and the degrees of freedom in the denominator are for the second sample. The first sample taken from the first population is defined as the sample that has the *larger* sample variance. The second sample taken from the second population is the sample with the *smaller* sample variance. Equation (10.9) defines the ***F* test for the ratio of two variances**.

F TEST STATISTIC FOR TESTING THE RATIO OF TWO VARIANCES

The F_{STAT} test statistic is equal to the variance of sample 1 (the larger sample variance) divided by the variance of sample 2 (the smaller sample variance).



Student Tip
Since the numerator of Equation (10.9) contains the larger variance, the F_{STAT} statistic is always greater than or equal to 1.0.

$$F_{STAT} = \frac{S_1^2}{S_2^2} \quad (10.9)$$

where

S_1^2 = variance of sample 1 (the larger sample variance)

S_2^2 = variance of sample 2 (the smaller sample variance)

n_1 = sample size selected from population 1

n_2 = sample size selected from population 2

$n_1 - 1$ = degrees of freedom from sample 1 (i.e., the numerator degrees of freedom)

$n_2 - 1$ = degrees of freedom from sample 2 (i.e., the denominator degrees of freedom)

The F_{STAT} test statistic follows an *F* distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

For a given level of significance, α , to test the null hypothesis of equality of population variances:

$$H_0: \sigma_1^2 = \sigma_2^2$$

against the alternative hypothesis that the two population variances are not equal:

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

you reject the null hypothesis if the computed F_{STAT} test statistic is greater than the upper-tail critical value, $F_{\alpha/2}$, from the *F* distribution, with $n_1 - 1$ degrees of freedom in the numerator and $n_2 - 1$ degrees of freedom in the denominator. Thus, the decision rule is

Reject H_0 if $F_{STAT} > F_{\alpha/2}$;
otherwise, do not reject H_0 .

To illustrate how to use the *F* test to determine whether the two variances are equal, return to the North Fork Beverages scenario on page 397 concerning the sales of the new cola in two different end-cap locations. To determine whether to use the pooled-variance *t* test or the separate-variance *t* test in Section 10.1, you can test the equality of the two population variances. The null and alternative hypotheses are

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Because you are defining sample 1 as the group with the larger sample variance, the rejection region in the upper tail of the F distribution contains $\alpha/2$. Using the level of significance $\alpha = 0.05$, the rejection region in the upper tail contains 0.025 of the distribution.

Because there are samples of 10 stores for each of the two end-cap locations, there are $10 - 1 = 9$ degrees of freedom in the numerator (the sample with the larger variance) and also in the denominator (the sample with the smaller variance). $F_{\alpha/2}$, the upper-tail critical value of the F distribution, is found directly from Table E.5, a portion of which is presented in Table 10.6. Because there are 9 degrees of freedom in the numerator and 9 degrees of freedom in the denominator, you find the upper-tail critical value, $F_{\alpha/2}$, by looking in the column labeled 9 and the row labeled 9. Thus, the upper-tail critical value of this F distribution is 4.03. Therefore, the decision rule is

Reject H_0 if $F_{STAT} > F_{0.025} = 4.03$;

otherwise, do not reject H_0 .

TABLE 10.6

Finding the Upper-Tail Critical Value of F with 9 and 9 Degrees of Freedom for an Upper-Tail Area of 0.025

Denominator df_2	Cumulative Probabilities = 0.975 Upper-Tail Area = 0.025 Numerator df_1						
	1	2	3	...	7	8	9
1	647.80	799.50	864.20	...	948.20	956.70	963.30
2	38.51	39.00	39.17	...	39.36	39.37	39.39
3	17.44	16.04	15.44	...	14.62	14.54	14.47
...	⋮	⋮	⋮	⋮	⋮	⋮	⋮
7	8.07	6.54	5.89	...	4.99	4.90	4.82
8	7.57	6.06	5.42	...	4.53	4.43	4.36
9	7.21	5.71	5.08	...	4.20	4.10	4.03

Source: Extracted from Table E.5.

Using Equation (10.9) on page 374 and the cola sales data (see Table 10.1 on page 349),

$$S_1^2 = (18.7264)^2 = 350.6778 \quad S_2^2 = (12.5433)^2 = 157.3333$$

so that

$$\begin{aligned} F_{STAT} &= \frac{S_1^2}{S_2^2} \\ &= \frac{350.6778}{157.3333} = 2.2289 \end{aligned}$$

Because $F_{STAT} = 2.2289 < 4.03$, you do not reject H_0 . Figure 10.14 shows the results for this test, including the p -value, 0.2482. Because $0.2482 > 0.05$, you conclude that there is no evidence of a significant difference in the variability of the sales of the new cola for the two end-cap locations.

In testing for a difference between two variances using the F test, you assume that each of the two populations is normally distributed. The F test is very sensitive to the normality assumption. If boxplots or normal probability plots suggest even a mild departure from normality for either of the two populations, you should not use the F test. If this happens, you should use the Levene test (see Section 11.1) or a nonparametric approach (see references 1 and 2).

In testing for the equality of variances as part of assessing the validity of the pooled-variance t test procedure, the F test is a two-tail test with $\alpha/2$ in the upper tail. However, when you are interested in examining the variability in situations other than the pooled-variance t test, the F test is often a one-tail test. Example 10.4 illustrates a one-tail test.

FIGURE 10.14

Excel and Minitab F test results for the two end-cap locations data

A		B
1 F Test for Differences in Two Variances		
2		
3 Data		
4 Level of Significance		0.05
5 Larger-Variance Sample		
6 Sample Size		10 =COUNT(DATACOPY!\$A:\$A)
7 Sample Variance		350.6778 =VAR.S(DATACOPY!\$A:\$A)
8 Smaller-Variance Sample		
9 Sample Size		10 =COUNT(DATACOPY!\$B:\$B)
10 Sample Variance		157.3333 =VAR.S(DATACOPY!\$B:\$B)
11		
12 Intermediate Calculations		
13 F Test Statistic		2.2289 =B7/B10
14 Population 1 Sample Degrees of Freedom		9 =B6 - 1
15 Population 2 Sample Degrees of Freedom		9 =B9 - 1
16		
17 Two-Tail Test		
18 Upper Critical Value		4.0260 =F.INV.RT(B4/2, B14, B15)
19 p-Value		0.2482 =2 * F.DIST.RT(B13, B14, B15)
20 Do not reject the null hypothesis		=IF(B19 < B4, "Reject the null hypothesis", "Do not reject the null hypothesis")

Alternative hypothesis	Sigma(Beverage) / Sigma(Produce) not = 1
Significance level	Alpha = 0.05
Statistics	
Variable N StDev Variance	
Beverage 10 18.726 350.678	
Produce 10 12.543 157.333	
Ratio of standard deviations = 1.493	
Ratio of variances = 2.229	
95% Confidence Intervals	
Distribution CI for StDev CI for Variance	
of Data Ratio Ratio	
Normal (0.744, 2.996) (0.554, 8.973)	
Continuous (0.664, 3.082) (0.441, 9.497)	
Tests	
Method DF1 DF2 Statistic P-Value	
F Test (normal) 9 9 2.23 0.248	
Levene's Test (any continuous) 1 18 1.27 0.275	

EXAMPLE 10.4**A One-Tail Test for the Difference Between Two Variances**

Waiting time is a critical issue at fast-food chains, which not only want to minimize the mean service time but also want to minimize the variation in the service time from customer to customer. One fast-food chain carried out a study to measure the variability in the waiting time (defined as the time in minutes from when an order was completed to when it was delivered to the customer) at lunch and breakfast at one of the chain's stores. The results were as follows:

$$\text{Lunch: } n_1 = 25 \quad S_1^2 = 4.4$$

$$\text{Breakfast: } n_2 = 21 \quad S_2^2 = 1.9$$

At the 0.05 level of significance, is there evidence that there is more variability in the service time at lunch than at breakfast? Assume that the population service times are normally distributed.

SOLUTION The null and alternative hypotheses are

$$H_0: \sigma_L^2 \leq \sigma_B^2$$

$$H_1: \sigma_L^2 > \sigma_B^2$$

The F_{STAT} test statistic is given by Equation (10.9) on page 374:

$$F_{STAT} = \frac{S_1^2}{S_2^2}$$

You use Table E.5 to find the upper critical value of the F distribution. With $n_1 - 1 = 25 - 1 = 24$ degrees of freedom in the numerator, $n_2 - 1 = 21 - 1 = 20$ degrees of freedom in the denominator, and $\alpha = 0.05$, the upper-tail critical value, $F_{0.05}$, is 2.08. The decision rule is

Reject H_0 if $F_{STAT} > 2.08$;

otherwise, do not reject H_0 .

From Equation (10.9) on page 374,

$$\begin{aligned}F_{STAT} &= \frac{S_1^2}{S_2^2} \\&= \frac{4.4}{1.9} = 2.3158\end{aligned}$$

Because $F_{STAT} = 2.3158 > 2.08$, you reject H_0 . Using a 0.05 level of significance, you conclude that there is evidence that there is more variability in the service time at lunch than at breakfast.

Problems for Section 10.4

LEARNING THE BASICS

10.36 Determine the upper-tail critical values of F in each of the following two-tail tests.

- a. $\alpha = 0.10, n_1 = 16, n_2 = 21$
- b. $\alpha = 0.05, n_1 = 16, n_2 = 21$
- c. $\alpha = 0.01, n_1 = 16, n_2 = 21$

10.37 Determine the upper-tail critical value of F in each of the following one-tail tests.

- a. $\alpha = 0.05, n_1 = 16, n_2 = 21$
- b. $\alpha = 0.01, n_1 = 16, n_2 = 21$

10.38 The following information is available for two samples selected from independent normally distributed populations:

$$\text{Population A: } n_1 = 25 \quad S_1^2 = 16$$

$$\text{Population B: } n_2 = 25 \quad S_2^2 = 25$$

- a. Which sample variance do you place in the numerator of F_{STAT} ?
- b. What is the value of F_{STAT} ?

10.39 The following information is available for two samples selected from independent normally distributed populations:

$$\text{Population A: } n_1 = 25 \quad S_1^2 = 161.9$$

$$\text{Population B: } n_2 = 25 \quad S_2^2 = 133.7$$

What is the value of F_{STAT} if you are testing the null hypothesis $H_0: \sigma_1^2 = \sigma_2^2$?

10.40 In Problem 10.39, how many degrees of freedom are there in the numerator and denominator of the F test?

10.41 In Problems 10.39 and 10.40, what is the upper-tail critical value for F if the level of significance, α , is 0.05 and the alternative hypothesis is $H_1: \sigma_1^2 \neq \sigma_2^2$?

10.42 In Problems 10.39 through 10.41, what is your statistical decision?

10.43 The following information is available for two samples selected from independent but very right-skewed populations:

$$\text{Population A: } n_1 = 16 \quad S_1^2 = 47.3$$

$$\text{Population B: } n_2 = 13 \quad S_2^2 = 36.4$$

Should you use the F test to test the null hypothesis of equality of variances? Discuss.

10.44 In Problem 10.43, assume that two samples are selected from independent normally distributed populations.

- a. At the 0.05 level of significance, is there evidence of a difference between σ_1^2 and σ_2^2 ?
- b. Suppose that you want to perform a one-tail test. At the 0.05 level of significance, what is the upper-tail critical value of F to determine whether there is evidence that $\sigma_1^2 > \sigma_2^2$? What is your statistical decision?

APPLYING THE CONCEPTS

10.45 A problem with a telephone line that prevents a customer from receiving or making calls is upsetting to both the customer and the telecommunications company. The file **Phone** contains samples of 20 problems reported to two different offices of a telecommunications company and the time to clear these problems (in minutes) from the customers' lines.

- a. At the 0.05 level of significance, is there evidence of a difference in the variability of the time to clear problems between the two central offices?
- b. Determine the p -value in (a) and interpret its meaning.
- c. What assumption do you need to make in (a) about the two populations in order to justify your use of the F test?
- d. Based on the results of (a) and (b), which t test defined in Section 10.1 should you use to compare the mean time to clear problems in the two central offices?

 **10.46** *Accounting Today* identified the top accounting firms in 10 geographic regions across the United States. Even though all 10 regions reported growth in 2012, the Southeast and Gulf Coast regions reported the highest combined growths, with 7.94% and 9.92%, respectively. A characteristic description of the accounting firms in the Southeast and Gulf Coast regions included the number of partners in the firm. The file **Accounting-Partners2** contains the number of partners. (Data extracted from bit.ly/11asPHm.)

- a. At the 0.05 level of significance, is there evidence of a difference in the variability in numbers of partners for Southeast region accounting firms and Gulf Coast accounting firms?
- b. Determine the p -value in (a) and interpret its meaning.
- c. What assumption do you have to make about the two populations in order to justify the use of the F test?

- d. Based on (a) and (b), which t test defined in Section 10.1 should you use to test whether there is a significant difference in the mean number of partners for Southeast region accounting firms and Gulf Coast accounting firms?

10.47 A bank with a branch located in a commercial district of a city has the business objective of improving the process for serving customers during the noon-to-1 P.M. lunch period. To do so, the waiting time (defined as the number of minutes that elapses from when the customer enters the line until he or she reaches the teller window) needs to be shortened to increase customer satisfaction. A random sample of 15 customers is selected and the waiting times are collected and stored in **Bank1**. These data are:

4.21	5.55	3.02	5.13	4.77	2.34	3.54	3.20
4.50	6.10	0.38	5.12	6.46	6.19	3.79	

Suppose that another branch, located in a residential area, is also concerned with the noon-to-1 P.M. lunch period. A random sample of 15 customers is selected and the waiting times are collected and stored in **Bank2**. These data are:

9.66	5.90	8.02	5.79	8.73	3.82	8.01	8.35
10.49	6.68	5.64	4.08	6.17	9.91	5.47	

- Is there evidence of a difference in the variability of the waiting time between the two branches? (Use $\alpha = 0.05$.)
- Determine the p -value in (a) and interpret its meaning.
- What assumption about the population distribution of each bank is necessary in (a)? Is the assumption valid for these data?
- Based on the results of (a), is it appropriate to use the pooled-variance t test to compare the means of the two branches?

10.48 An important feature of digital cameras is battery life, the number of shots that can be taken before the battery needs to be recharged. The file **Cameras** contains the battery life of 11

subcompact cameras and 7 compact cameras. (Data extracted from "Cameras," *Consumer Reports*, July 2012, pp. 42–44.)

- Is there evidence of a difference in the variability of the battery life between the two types of digital cameras? (Use $\alpha = 0.05$.)
- Determine the p -value in (a) and interpret its meaning.
- What assumption about the population distribution of the two types of cameras is necessary in (a)? Is the assumption valid for these data?
- Based on the results of (a), which t test defined in Section 10.1 should you use to compare the mean battery life of the two types of cameras?

10.49 An article appearing in *The Exponent*, an independent college newspaper published by the Purdue Student Publishing Foundation, reported that the average American college student spends 1 hour (60 minutes) on Facebook daily. (Data extracted from bit.ly/NQRCJQ.) You wonder if there is a difference between males and females. You select a sample of 60 Facebook users (30 males and 30 females) at your college and collect data about the time spent on Facebook per day (in minutes) and store these data in **FacebookTime2**.

- Using a 0.05 level of significance, is there evidence of a difference in the variances of time spent on Facebook per day between males and females?
- On the basis of the results in (a), which t test defined in Section 10.1 should you use to compare the means of males and females? Discuss.

10.50 Is there a difference in the variation of the yield of five-year certificates of deposit (CDs) in different cities? The file **FiveYearCDRate** contains the yields for a five-year CD for ten banks in New York and eight banks in Los Angeles, as of June 3, 2013. (Data extracted from www.Bankrate.com, June 3, 2013.) At the 0.05 level of significance, is there evidence of a difference in the variance of the yield of five-year CDs in the two cities? Assume that the population yields are normally distributed.

USING STATISTICS

For North Fork, Are There Different Means to the Ends? Revisited

In the North Fork Beverages scenario, you were a regional sales manager for North Fork Beverages. You compared the sales volume of your new HandMade Citrus Cola when the product was featured in the beverage aisle end-cap to the sales volume when the product was featured in the end-cap by the produce department. An experiment was performed in which 10 stores used the beverage end-cap location and 10 stores used the produce end-cap location. Using a t test for the difference between two means, you were able to conclude that the mean sales using the produce end-cap location are higher than the mean sales for the beverage end-cap location. A confidence interval allowed you to infer with 95% confidence that population mean amount sold at the produce

end-cap location was between 6.73 and 36.67



Fotolia

cases more than the beverage end-cap location. You also performed the F test for the difference between two variances to see if the store-to-store variability in sales in stores using the produce end-cap location differed from the store-to-store variability in sales in stores using the beverage end-cap location. You concluded that there was no significant difference in the variability of the sales of cola for the two display locations. As a regional sales manager, you decide to lease the produce end-cap location in all FoodPlace Supermarkets during your next sales promotional period.

SUMMARY

In this chapter, you were introduced to a variety of tests for two samples. For situations in which the samples are independent, you learned statistical test procedures for analyzing possible differences between means, proportions, and variances. In addition, you learned a test procedure that is frequently used when analyzing differences between the means of two related samples. Remember that you need to select the test that is most appropriate for a given set of conditions and to critically investigate the validity of the assumptions underlying each of the hypothesis-testing procedures.

Table 10.7 provides a list of topics covered in this chapter. The roadmap in Figure 10.15 illustrates the steps needed in determining which two-sample test of hypothesis to use. The following are the questions you need to consider:

1. What type of variables do you have? If you are dealing with categorical variables, use the Z test for the difference

between two proportions. (This test assumes independent samples.)

2. If you have a numerical variable, determine whether you have independent samples or related samples. If you have related samples, and you can assume approximate normality, use the paired *t* test.
3. If you have independent samples, is your focus on variability or central tendency? If the focus is on variability, and you can assume approximate normality, use the *F* test.
4. If your focus is central tendency and you can assume approximate normality, determine whether you can assume that the variances of the two populations are equal. (This assumption can be tested using the *F* test.)
5. If you can assume that the two populations have equal variances, use the pooled-variance *t* test. If you cannot assume that the two populations have equal variances, use the separate-variance *t* test.

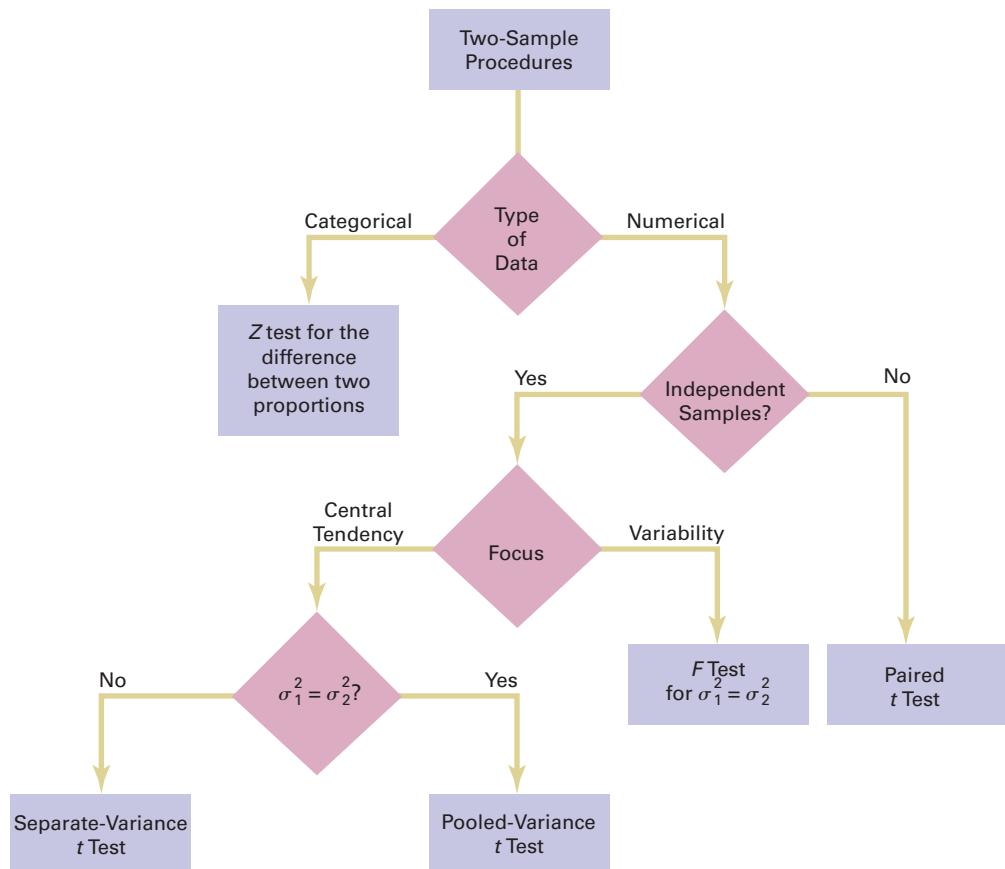
TABLE 10.7

Summary of Topics in Chapter 10

TYPE OF ANALYSIS	TYPES OF DATA	
	Numerical	Categorical
Comparing two populations	<i>t</i> tests for the difference in the means of two independent populations (Section 10.1) Paired <i>t</i> test (Section 10.2) <i>F</i> test for the difference between two variances (Section 10.4)	Z test for the difference between two proportions (Section 10.3)

FIGURE 10.15

Roadmap for selecting a test of hypothesis for two samples



REFERENCES

1. Conover, W. J. *Practical Nonparametric Statistics*, 3rd ed. New York: Wiley, 2000.
2. Daniel, W. *Applied Nonparametric Statistics*, 2nd ed. Boston: Houghton Mifflin, 1990.
3. Microsoft Excel 2013. Redmond, WA: Microsoft Corp., 2012.
4. Minitab Release 16 State College, PA: Minitab, 2010.
5. Satterthwaite, F. E. "An Approximate Distribution of Estimates of Variance Components." *Biometrics Bulletin*, 2(1946): 110–114.
6. Snedecor, G. W., and W. G. Cochran. *Statistical Methods*, 8th ed. Ames, IA: Iowa State University Press, 1989.

KEY EQUATIONS

Pooled-Variance t Test for the Difference Between Two Means

$$t_{STAT} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (10.1)$$

Confidence Interval Estimate for the Difference Between the Means of Two Independent Populations

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (10.2)$$

or

$$\begin{aligned} (\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} &\leq \mu_1 - \mu_2 \\ &\leq (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \end{aligned}$$

Separate-Variance t Test for the Difference Between Two Means

$$t_{STAT} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (10.3)$$

Computing Degrees of Freedom in the Separate-Variance t Test

$$V = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{\left(\frac{S_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2} \right)^2}{n_2 - 1}} \quad (10.4)$$

Paired t Test for the Mean Difference

$$t_{STAT} = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}} \quad (10.5)$$

Confidence Interval Estimate for the Mean Difference

$$\bar{D} \pm t_{\alpha/2} \frac{S_D}{\sqrt{n}} \quad (10.6)$$

or

$$\bar{D} - t_{\alpha/2} \frac{S_D}{\sqrt{n}} \leq \mu_D \leq \bar{D} + t_{\alpha/2} \frac{S_D}{\sqrt{n}}$$

Z Test for the Difference Between Two Proportions

$$Z_{STAT} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1 - \bar{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (10.7)$$

Confidence Interval Estimate for the Difference Between Two Proportions

$$(p_1 - p_2) \pm Z_{\alpha/2} \sqrt{\left(\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2} \right)} \quad (10.8)$$

or

$$\begin{aligned} (p_1 - p_2) - Z_{\alpha/2} \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} &\leq (\pi_1 - \pi_2) \\ &\leq (p_1 - p_2) + Z_{\alpha/2} \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} \end{aligned}$$

F Test Statistic for Testing the Ratio of Two Variances

$$F_{STAT} = \frac{S_1^2}{S_2^2} \quad (10.9)$$

KEY TERMS

F distribution 374
F test for the ratio of two variances 374
 matched samples 359
 paired *t* test for the mean difference 360

pooled-variance *t* test 348
 repeated measurements 359
 robust 351
 separate-variance *t* test 354

two-sample tests 348
Z test for the difference between two proportions 367

CHECKING YOUR UNDERSTANDING

10.51 What are some of the criteria used in the selection of a particular hypothesis-testing procedure?

10.52 Under what conditions should you use the pooled-variance *t* test to examine possible differences in the means of two independent populations?

10.53 Under what conditions should you use the *F* test to examine possible differences in the variances of two independent populations?

10.54 What is the distinction between two independent populations and two related populations?

10.55 What is the distinction between repeated measurements and matched items?

10.56 When you have two independent populations, explain the similarities and differences between the test of hypothesis for the difference between the means and the confidence interval estimate for the difference between the means.

10.57 Under what conditions should you use the paired *t* test for the mean difference between two related populations?

CHAPTER REVIEW PROBLEMS

10.58 The American Society for Quality (ASQ) conducted a salary survey of all its members. ASQ members work in all areas of manufacturing and service-related institutions, with a common theme of an interest in quality. Two job titles are black belt and green belt. (See Section 19.6 for a description of these titles in a Six Sigma quality improvement initiative.) Descriptive statistics concerning salaries for these two job titles are given in the following table:

Job Title	Sample Size	Mean	Standard Deviation
Black belt	121	93,946	21,166
Green belt	26	74,173	23,399

Source: Data extracted from "QP Salary Survey," *Quality Progress*, December 2012, p. 29.

business students each year. The file **StudyTime** contains the gender and the number of hours spent studying in a typical week for the sampled students.

- a. At the 0.05 level of significance, is there a difference in the variance of the study time for male students and female students?
- b. Using the results of (a), which *t* test is appropriate for comparing the mean study time for male and female students?
- c. At the 0.05 level of significance, conduct the test selected in (b).
- d. Write a short summary of your findings.

10.60 Do males and females differ in the amount of time they talk on the phone and the number of text messages they send? A study reported that women spent a mean of 818 minutes per month talking as compared to 716 minutes per month for men. (Data extracted from "Women Talk and Text More," *USA Today*, February 1, 2011, p. 1A.) The sample sizes were not reported. Suppose that the sample sizes were 100 each for women and men and that the standard deviation for women was 125 minutes per month as compared to 100 minutes per month for men.

- a. Using a 0.01 level of significance, is there evidence of a difference in the variances of the amount of time spent talking between women and men?
- b. To test for a difference in the mean talking time of women and men, is it most appropriate to use the pooled-variance *t* test or the separate-variance *t* test? Use the most appropriate test to determine if there is a difference in the amount of time spent talking between women and men.

- a. Using a 0.05 level of significance, is there a difference in the variability of salaries between black belts and green belts?
- b. Based on the result of (a), which *t* test defined in Section 10.1 is appropriate for comparing mean salaries?
- c. Using a 0.05 level of significance, is the mean salary of black belts greater than the mean salary of green belts?

10.59 Do male and female students study the same amount per week? In a recent year, 58 sophomore business students were surveyed at a large university that has more than 1,000 sophomore

- The article also reported that women sent a mean of 716 text messages per month compared to 555 per month for men. Suppose that the standard deviation for women was 150 text messages per month compared to 125 text messages per month for men.
- Using a 0.01 level of significance, is there evidence of a difference in the variances of the number of text messages sent per month by women and men?
 - Based on the results of (c), use the most appropriate test to determine, at the 0.01 level of significance, whether there is evidence of a difference in the mean number of text messages sent per month by women and men.

10.61 The file **Restaurants** contains the ratings for food, décor, service, and the price per person for a sample of 50 restaurants located in a city and 50 restaurants located in a suburb. Completely analyze the differences between city and suburban restaurants for the variables food rating, décor rating, service rating, and cost per person, using $\alpha = 0.05$.

Source: Data extracted from *Zagat Survey 2013 New York City Restaurants* and *Zagat Survey 2012–2013 Long Island Restaurants*.

10.62 A computer information systems professor is interested in studying the amount of time it takes students enrolled in the Introduction to Computers course to write a program in VB.NET. The professor hires you to analyze the following results (in minutes), stored in **VB**, from a random sample of nine students:

10 13 9 15 12 13 11 13 12

- At the 0.05 level of significance, is there evidence that the population mean time is greater than 10 minutes? What will you tell the professor?
- Suppose that the professor, when checking her results, realizes that the fourth student needed 51 minutes rather than the recorded 15 minutes to write the VB.NET program. At the 0.05 level of significance, reanalyze the question posed in (a), using the revised data. What will you tell the professor now?
- The professor is perplexed by these paradoxical results and requests an explanation from you regarding the justification for the difference in your findings in (a) and (b). Discuss.
- A few days later, the professor calls to tell you that the dilemma is completely resolved. The original number 15 (the fourth data value) was correct, and therefore your findings in (a) are being used in the article she is writing for a computer journal. Now she wants to hire you to compare the results from that group of Introduction to Computers students against those from a sample of 11 computer majors in order to determine whether there is evidence that computer majors can write a VB.NET program in less time than introductory students. For the computer majors, the sample mean is 8.5 minutes, and the sample standard deviation is 2.0 minutes. At the 0.05 level of significance, completely analyze these data. What will you tell the professor?
- A few days later, the professor calls again to tell you that a reviewer of her article wants her to include the p -value for the “correct” result in (a). In addition, the professor inquires about an unequal-variances problem, which the reviewer wants her to discuss in her article. In your own words, discuss the concept of p -value and also describe the unequal-variances problem. Then, determine the p -value in (a) and discuss whether the unequal-variances problem had any meaning in the professor’s study.

10.63 Do Pinterest shoppers and Facebook shoppers differ with respect to spending behavior? A study of browser-based shopping sessions reported that Pinterest shoppers spent a mean of \$153 per order and Facebook shoppers spent a mean of \$85 per order. (Data extracted from bitly/14wG1YI.) Suppose that the study consisted of 500 Pinterest shoppers and 500 Facebook shoppers, and the standard deviation of the order value was \$150 for Pinterest shoppers and \$80 for Facebook shoppers. Assume a level of significance of 0.05.

- Is there evidence of a difference in the variances of the order values between Pinterest shoppers and Facebook shoppers?
- Is there evidence of a difference in the mean order value between Pinterest shoppers and Facebook shoppers?
- Construct a 95% confidence interval estimate for the difference in mean order value between Pinterest shoppers and Facebook shoppers.

10.64 The lengths of life (in hours) of a sample of 40 20-watt compact fluorescent light bulbs produced by manufacturer A and a sample of 40 20-watt compact fluorescent light bulbs produced by manufacturer B are stored in **Bulbs**. Completely analyze the differences between the lengths of life of the compact fluorescent light bulbs produced by the two manufacturers. (Use $\alpha = 0.05$.)

10.65 A hotel manager looks to enhance the initial impressions that hotel guests have when they check in. Contributing to initial impressions is the time it takes to deliver a guest’s luggage to the room after check-in. A random sample of 20 deliveries on a particular day were selected in Wing A of the hotel, and a random sample of 20 deliveries were selected in Wing B. The results are stored in **Luggage**. Analyze the data and determine whether there is a difference between the mean delivery times in the two wings of the hotel. (Use $\alpha = 0.05$.)

10.66 The owner of a restaurant that serves Continental-style entrées has the business objective of learning more about the patterns of patron demand during the Friday-to-Sunday weekend time period. She decided to study the demand for dessert during this time period. In addition to studying whether a dessert was ordered, she will study the gender of the individual and whether a beef entrée was ordered. Data were collected from 630 customers and organized in the following contingency tables:

		GENDER		
		Male	Female	Total
DESSERT ORDERED	Yes	50	96	146
	No	250	234	484
	Total	300	330	630
BEEF ENTRÉE				
DESSERT ORDERED		Yes	No	Total
DESSERT ORDERED	Yes	74	68	142
	No	123	365	488
	Total	197	433	630

- At the 0.05 level of significance, is there evidence of a difference between males and females in the proportion who order dessert?
- At the 0.05 level of significance, is there evidence of a difference in the proportion who order dessert based on whether a beef entrée has been ordered?

10.67 The manufacturer of Boston and Vermont asphalt shingles knows that product weight is a major factor in the customer's perception of quality. Moreover, the weight represents the amount of raw materials being used and is therefore very important to the company from a cost standpoint. The last stage of the assembly line packages the shingles before they are placed on wooden pallets. Once a pallet is full (a pallet for most brands holds 16 squares of shingles), it is weighed, and the measurement is recorded. The file **Pallet** contains the weight (in pounds) from a sample of 368 pallets of Boston shingles and 330 pallets of Vermont shingles. Completely analyze the differences in the weights of the Boston and Vermont shingles, using $\alpha = 0.05$.

10.68 The manufacturer of Boston and Vermont asphalt shingles provides its customers with a 20-year warranty on most of its products. To determine whether a shingle will last as long as the warranty period, the manufacturer conducts accelerated-life testing. Accelerated-life testing exposes the shingle to the stresses it would be subject to in a lifetime of normal use in a laboratory setting via an experiment that takes only a few minutes to conduct. In this test, a shingle is repeatedly scraped with a brush for a short period of time, and the shingle granules removed by the brushing are weighed (in grams). Shingles that experience low amounts of

granule loss are expected to last longer in normal use than shingles that experience high amounts of granule loss. In this situation, a shingle should experience no more than 0.8 grams of granule loss if it is expected to last the length of the warranty period. The file **Granule** contains a sample of 170 measurements made on the company's Boston shingles and 140 measurements made on Vermont shingles. Completely analyze the differences in the granule loss of the Boston and Vermont shingles, using $\alpha = 0.05$.

10.69 There are a very large number of mutual funds from which an investor can choose. Each mutual fund has its own mix of different types of investments. The data in **BestFunds1** present the one-year return and the three-year annualized return for the 10 best short-term bond funds and the 10 best long-term bond funds, according to the *U.S. News & World Report*. (Data extracted from money.usnews.com/mutual-funds/rankings.) Analyze the data and determine whether any differences exist between short-term and long-term bond funds. (Use the 0.05 level of significance.)

REPORT WRITING EXERCISE

10.70 Referring to the results of Problems 10.67 and 10.68 concerning the weight and granule loss of Boston and Vermont shingles, write a report that summarizes your conclusions.

CASES FOR CHAPTER 10

Managing Ashland MultiComm Services

AMS communicates with customers who subscribe to cable television services through a special secured email system that sends messages about service changes, new features, and billing information to in-home digital set-top boxes for later display. To enhance customer service, the operations department established the business objective of reducing the amount of time to fully update each subscriber's

set of messages. The department selected two candidate messaging systems and conducted an experiment in which 30 randomly chosen cable subscribers were assigned one of the two systems (15 assigned to each system). Update times were measured, and the results are organized in Table AMS10.1 and stored in **AMS10**.

TABLE AMS 10.1

Update Times (in seconds) for Two Different Email Interfaces

	Email Interface 1	Email Interface 2
	4.13	3.71
	3.75	3.89
	3.93	4.22
	3.74	4.57
	3.36	4.24
	3.85	3.90
	3.26	4.09
	3.73	4.05
	4.06	4.07
	3.33	3.80
	3.96	4.36
	3.57	4.38
	3.13	3.49
	3.68	3.57
	3.63	4.74

- Analyze the data in Table AMS10.1 and write a report to the computer operations department that indicates your findings. Include an appendix in which you discuss the reason you selected a particular statistical test to compare the two independent groups of callers.
- Suppose that instead of the research design described in the case, there were only 15 subscribers sampled, and the

update process for each subscriber email was measured for each of the two messaging systems. Suppose that the results were organized in Table AMS10.1—making each row in the table a pair of values for an individual subscriber. Using these suppositions, reanalyze the Table AMS10.1 data and write a report for presentation to the team that indicates your findings.

Digital Case

Apply your knowledge about hypothesis testing in this Digital Case, which continues the cereal-fill packaging dispute Digital Case from Chapters 7 and 9.

Even after the recent public experiment about cereal box weights, Consumers Concerned About Cereal Cheaters (CCACC) remains convinced that Oxford Cereals has misled the public. The group has created and circulated **More-Cheating.pdf**, a document in which it claims that cereal

boxes produced at Plant Number 2 in Springville weigh less than the claimed mean of 368 grams. Review this document and then answer the following questions:

- Do the CCACC's results prove that there is a statistically significant difference in the mean weights of cereal boxes produced at Plant Numbers 1 and 2?
- Perform the appropriate analysis to test the CCACC's hypothesis. What conclusions can you reach based on the data?

Sure Value Convenience Stores

You work in the corporate office for a nationwide convenience store franchise that operates nearly 10,000 stores. The per-store daily customer count (i.e., the mean number of customers in a store in one day) has been steady, at 900, for some time. To increase the customer count, the chain is considering cutting prices for coffee beverages. The small size will now be either \$0.59 or \$0.79 instead of \$0.99. Even with this reduction in price, the chain will have a 40% gross margin on coffee.

The question to be determined is how much to cut prices to increase the daily customer count without reducing the gross margin on coffee sales too much. The chain decides to carry out an experiment in a sample of 30 stores where customer counts have been running almost exactly at the national average of 900. In 15 of the stores, the price of a small coffee will now be \$0.59 instead of \$0.99, and in 15 other stores, the price of a small coffee will now be \$0.79. After four weeks, the 15 stores that priced the small

coffee at \$0.59 had a mean daily customer count of 964 and a standard deviation of 88, and the 15 stores that priced the small coffee at \$0.79 had a mean daily customer count of 941 and a standard deviation of 76. Analyze these data (using the 0.05 level of significance) and answer the following questions.

- Does reducing the price of a small coffee to either \$0.59 or \$0.79 increase the mean per-store daily customer count?
- If reducing the price of a small coffee to either \$0.59 or \$0.79 increases the mean per-store daily customer count, is there any difference in the mean per-store daily customer count between stores in which a small coffee was priced at \$0.59 and stores in which a small coffee was priced at \$0.79?
- What price do you recommend for a small coffee?

CardioGood Fitness

Return to the CardioGood Fitness case first presented on page 83. Using the data stored in **CardioGood Fitness**:

- Determine whether differences exist between males and females in their age in years, education in years, annual household income (\$), mean number of times the

customer plans to use the treadmill each week, and mean number of miles the customer expects to walk or run each week.

- Write a report to be presented to the management of CardioGood Fitness detailing your findings.

More Descriptive Choices Follow-Up

Follow up the Using Statistics scenario “More Descriptive Choices, Revisited” on page 138 by determining whether there is a difference in the 3-year return percentage, 5-year

return percentages, and 10-year return percentages of the growth and value funds (stored in [Retirement Funds](#)).

Clear Mountain State Student Surveys

1. The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. It creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in [UndergradSurvey](#)).
 - a. At the 0.05 level of significance, is there evidence of a difference between males and females in grade point average, expected starting salary, number of social networking sites registered for, age, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?
 - b. At the 0.05 level of significance, is there evidence of a difference between students who plan to go to graduate school and those who do not plan to go to graduate school in grade point average, expected starting salary, number of social networking sites registered for, age, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?
2. The dean of students at CMSU has learned about the undergraduate survey and has decided to undertake a similar survey for graduate students at Clear Mountain State. She creates and distributes a survey of 14 questions and receives responses from 44 graduate students (stored in [GradSurvey](#)). For these data, at the 0.05 level of significance, is there evidence of a difference between males and females in age, undergraduate grade point average, graduate grade point average, expected salary upon graduation, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?

CHAPTER 10 EXCEL GUIDE

EG10.1 COMPARING the MEANS of TWO INDEPENDENT POPULATIONS

Pooled-Variance t Test for the Difference Between Two Means

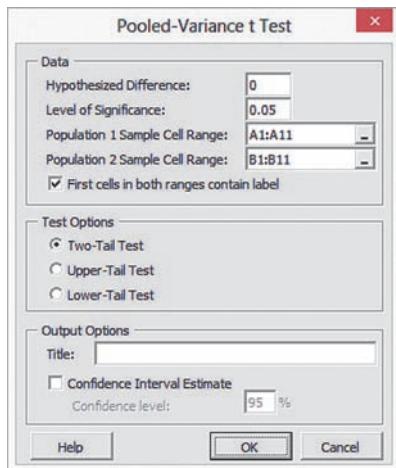
Key Technique Use the **T.INV.2T**(*level of significance, degrees of freedom*) function to compute the lower and upper critical values and use the **T.DIST.2T**(*absolute value of the t test statistic, degrees of freedom*) to compute the *p*-value.

Example Perform the Figure 10.3 pooled-variance *t* test for the two end-cap locations data shown on page 350.

PHStat Use Pooled-Variance t Test.

For the example, open to the **DATA worksheet** of the **Cola workbook**. Select **PHStat** → **Two-Sample Tests (Unsummarized Data)** → **Pooled-Variance t Test**. In the procedure's dialog box (shown below):

1. Enter **0** as the **Hypothesized Difference**.
2. Enter **0.05** as the **Level of Significance**.
3. Enter **A1:A11** as the **Population 1 Sample Cell Range**.
4. Enter **B1:B11** as the **Population 2 Sample Cell Range**.
5. Check **First cells in both ranges contain label**.
6. Click **Two-Tail Test**.
7. Enter a **Title** and click **OK**.



When using summarized data, select **PHStat** → **Two-Sample Tests (Summarized Data)** → **Pooled-Variance t Test**. In that procedure's dialog box, enter the hypothesized difference and level of significance, as well as the sample size, sample mean, and sample standard deviation for each sample.

In-Depth Excel Use the **COMPUTE worksheet** of the **Pooled-Variance T workbook** as a template.

The worksheet already contains the data and formulas to use the unsummarized data for the example. For other problems, use this worksheet with either unsummarized or summarized data. For unsummarized data, paste the data in columns A and B in the **DATA-COPY worksheet** and keep the COMPUTE worksheet formulas that compute the sample size, sample mean, and sample standard deviation in the cell range B7:B13. For summarized data, replace the formulas in the cell range B7:B13 with the sample statistics and ignore the DATACOPY worksheet.

Use the **COMPUTE_LOWER** or **COMPUTE_UPPER worksheets** in the same workbook as templates for performing one-tail pooled-variance *t* tests with either unsummarized or summarized data. If you use an Excel version older than Excel 2010, use the **COMPUTE_Older** worksheet as a template for both the two-tail and one-tail tests.

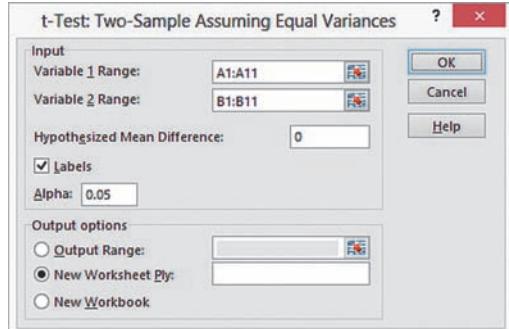
Analysis ToolPak Use t-Test: Two-Sample Assuming Equal Variances.

For the example, open to the **DATA worksheet** of the **Cola workbook** and:

1. Select **Data** → **Data Analysis**.
2. In the Data Analysis dialog box, select **t-Test: Two-Sample Assuming Equal Variances** from the **Analysis Tools** list and then click **OK**.

In the procedure's dialog box (shown below):

3. Enter **A1:A11** as the **Variable 1 Range**.
4. Enter **B1:B11** as the **Variable 2 Range**.
5. Enter **0** as the **Hypothesized Mean Difference**.
6. Check **Labels** and enter **0.05** as **Alpha**.
7. Click **New Worksheet Ply**.
8. Click **OK**.



Results (shown below) appear in a new worksheet that contains both two-tail and one-tail test critical values and p -values. Unlike the results shown in Figure 10.3, only the positive (upper) critical value is listed for the two-tail test.

	A	B	C
1	t-Test: Two-Sample Assuming Equal Variances		
3	Beverage		Produce
4	Mean	50.3	72
5	Variance	350.6778	157.3333
6	Observations	10	10
7	Pooled Variance	254.0056	
8	Hypothesized Mean Difference	0	
9	df	18	
10	t Stat	-3.0446	
11	P(T<=t) one-tail	0.0035	
12	t Critical one-tail	1.7341	
13	P(T<=t) two-tail	0.0070	
14	t Critical two-tail	2.1009	

Confidence Interval Estimate for the Difference Between Two Means

PHStat Modify the *PHStat* instructions for the pooled-variance t test. In step 7, check **Confidence Interval Estimate** and enter a **Confidence Level** in its box, in addition to entering a **Title** and clicking **OK**.

In-Depth Excel Use the *In-Depth Excel* instructions for the pooled-variance t test. The Pooled-Variance T workbook worksheets include a confidence interval estimate for the difference between two means in the cell range D3:E16.

t Test for the Difference Between Two Means, Assuming Unequal Variances

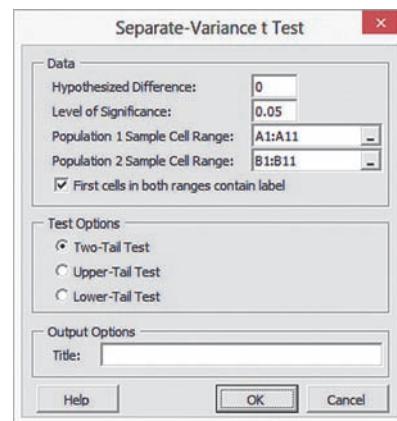
Key Technique Use the **T.INV.2T(*level of significance, degrees of freedom*)** function to compute the lower and upper critical values and use the **T.DIST.2T(*absolute value of the t test statistic, degrees of freedom*)** to compute the p -value.

Example Perform the Figure 10.7 separate-variance t test for the two end-cap locations data shown on page 356.

PHStat Use Separate-Variance t Test.

For the example, open to the **DATA worksheet** of the **Cola workbook**. Select **PHStat → Two-Sample Tests (Unsummarized Data) → Separate-Variance t Test**. In the procedure's dialog box (shown in the right column):

1. Enter **0** as the **Hypothesized Difference**.
2. Enter **0.05** as the **Level of Significance**.
3. Enter **A1:A11** as the **Population 1 Sample Cell Range**.
4. Enter **B1:B11** as the **Population 2 Sample Cell Range**.
5. Check **First cells in both ranges contain label**.
6. Click **Two-Tail Test**.
7. Enter a **Title** and click **OK**.



When using summarized data, select **PHStat → Two-Sample Tests (Summarized Data) → Separate-Variance t Test**. In that procedure's dialog box, enter the hypothesized difference and the level of significance, as well as the sample size, sample mean, and sample standard deviation for each group.

In-Depth Excel Use the **COMPUTE worksheet** of the **Separate-Variance T workbook** as a template.

The worksheet already contains the data and formulas to use the unsummarized data for the example. For other problems, use the COMPUTE worksheet with either unsummarized or summarized data. For unsummarized data, paste the data in columns A and B in the **DATACOPY worksheet** and keep the COMPUTE worksheet formulas that compute the sample size, sample mean, and sample standard deviation in the cell range B7:B13. For summarized data, replace those formulas in the cell range B7:B13 with the sample statistics and ignore the DATACOPY worksheet.

Use the **COMPUTE_LOWER** or **COMPUTE_UPPER** worksheets in the same workbook as templates for performing one-tail pooled-variance t tests with either unsummarized or summarized data. If you use an Excel version older than Excel 2010, use the **COMPUTE_Older** worksheet as a template for both the two-tail and one-tail tests.

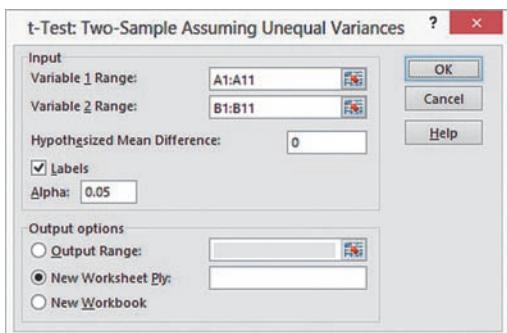
Analysis ToolPak Use **t-Test: Two-Sample Assuming Unequal Variances**.

For the example, open to the **DATA worksheet** of the **Cola workbook** and:

1. Select **Data → Data Analysis**.
2. In the Data Analysis dialog box, select **t-Test: Two-Sample Assuming Unequal Variances** from the **Analysis Tools** list and then click **OK**.

In the procedure's dialog box (shown on page 388):

3. Enter **A1:A11** as the **Variable 1 Range**.
4. Enter **B1:B11** as the **Variable 2 Range**.
5. Enter **0** as the **Hypothesized Mean Difference**.
6. Check **Labels** and enter **0.05** as **Alpha**.
7. Click **New Worksheet Ply**.
8. Click **OK**.



Results (shown below) appear in a new worksheet that contains both two-tail and one-tail test critical values and p -values. Unlike the results shown in Figure 10.7, only the positive (upper) critical value is listed for the two-tail test. Because the Analysis ToolPak uses table lookups to approximate the critical values and the p -value, the results will differ slightly from the values shown in Figure 10.7.

	A	B	C
1 t-Test: Two-Sample Assuming Unequal Variances			
2			
3		Beverage	Produce
4 Mean		50.3	72
5 Variance		350.6778	157.3333
6 Observations		10	10
7 Hypothesized Mean Difference		0	
8 df		16	
9 t Stat		-3.04455	
10 P(T<=t) one-tail		0.003863	
11 t Critical one-tail		1.745884	
12 P(T>=t) two-tail		0.007726	
13 t Critical two-tail		2.119905	

EG10.2 COMPARING the MEANS of TWO RELATED POPULATIONS

Paired t Test

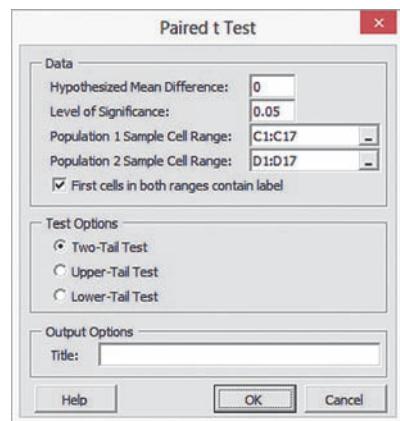
Key Technique Use the **T.INV.2T**(*level of significance, degrees of freedom*) function to compute the lower and upper critical values and use the **T.DIST.2T**(*absolute value of the t test statistic, degrees of freedom*) to compute the p -value.

Example Perform the Figure 10.9 paired t test for the textbook price data shown on page 363.

PHStat Use Paired t Test.

For the example, open to the **DATA worksheet** of the **BookPrices workbook**. Select **PHStat** → **Two-Sample Tests (Unsummarized Data)** → **Paired t Test**. In the procedure's dialog box (shown in the right column):

1. Enter **0** as the **Hypothesized Mean Difference**.
2. Enter **0.05** as the **Level of Significance**.
3. Enter **C1:C17** as the **Population 1 Sample Cell Range**.
4. Enter **D1:D17** as the **Population 2 Sample Cell Range**.
5. Check **First cells in both ranges contain label**.
6. Click **Two-Tail Test**.
7. Enter a **Title** and click **OK**.



The procedure creates two worksheets, one of which is similar to the PtCalcs worksheet discussed in the following *In-Depth Excel* section. When using summarized data, select **PHStat** → **Two-Sample Tests (Summarized Data)** → **Paired t Test**. In that procedure's dialog box, enter the hypothesized mean difference, the level of significance, and the differences cell range.

In-Depth Excel Use the **COMPUTE** and **PtCalcs worksheets** of the **Paired T workbook** as a template.

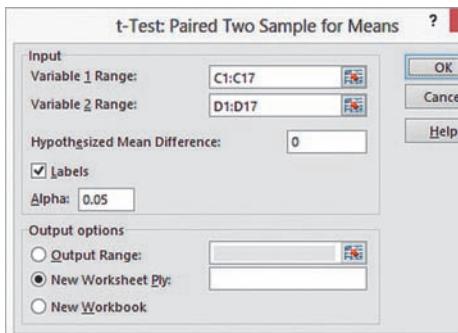
The COMPUTE and supporting PtCalcs worksheets already contain the textbook price data for the example. The PtCalcs worksheet also computes the differences that allow the COMPUTE worksheet to compute the S_D in cell B11.

For other problems, paste the unsummarized data into columns A and B of the PtCalcs worksheet. For sample sizes greater than 16, select cell C17 and copy the formula in that cell down through the last data row. For sample sizes less than 16, delete the column C formulas for which there are no column A and B values. If you know the sample size, \bar{D} , and S_D values, you can ignore the PtCalcs worksheet and enter the values in cells B8, B9, and B11 of the COMPUTE worksheet, overwriting the formulas that those cells contain.

Use the similar **COMPUTE_LOWER** and **COMPUTE_UPPER** worksheets in the same workbook as templates for performing one-tail tests. If you use an Excel version older than Excel 2010, use the **COMPUTE_OLDER** worksheet as a template for both the two-tail and one-tail tests.

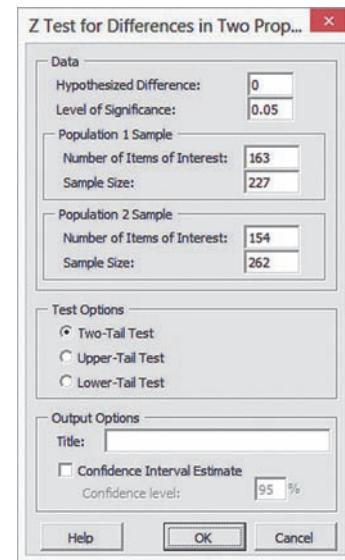
Analysis ToolPak Use **t-Test: Paired Two Sample for Means**. For the example, open to the **DATA worksheet** of the **BookPrices workbook** and:

1. Select **Data** → **Data Analysis**.
 2. In the Data Analysis dialog box, select **t-Test: Paired Two Sample for Means** from the **Analysis Tools** list and then click **OK**.
- In the procedure's dialog box (shown on page 389):
3. Enter **C1:C17** as the **Variable 1 Range**.
 4. Enter **D1:D17** as the **Variable 2 Range**.
 5. Enter **0** as the **Hypothesized Mean Difference**.
 6. Check **Labels** and enter **0.05** as **Alpha**.
 7. Click **New Worksheet Ply**.
 8. Click **OK**.



Results (shown below) appear in a new worksheet that contains both two-tail and one-tail test critical values and p -values. Unlike in Figure 10.9, only the positive (upper) critical value is listed for the two-tail test.

	A	B	C
1 t-Test: Paired Two Sample for Means			
2			
3		Bookstore	Online
4 Mean	153.6344	111.0331	
5 Variance	7913.0962	3594.7263	
6 Observations	16	16	
7 Pearson Correlation	0.8990		
8 Hypothesized Mean Difference	0		
9 df	15		
10 t Stat	3.8908		
11 P(T<=t) one-tail	0.0007		
12 t Critical one-tail	1.7531		
13 P(T<=t) two-tail	0.0014		
14 t Critical two-tail	2.1314		



EG10.3 COMPARING the PROPORTIONS of TWO INDEPENDENT POPULATIONS

Z Test for the Difference Between Two Proportions

Key Technique Use the **NORM.S.INV** (*percentage*) function to compute the critical values and use the **NORM.S.DIST** (*absolute value of the Z test statistic, True*) function to compute the p -value.

Example Perform the Figure 10.13 Z test for the hotel guest satisfaction survey shown on page 370.

PHStat Use Z Test for Differences in Two Proportions.

For the example, select **PHStat → Two-Sample Tests (Summarized Data) → Z Test for Differences in Two Proportions**. In the procedure's dialog box (shown in the right column):

- Enter 0 as the **Hypothesized Difference**.
- Enter 0.05 as the **Level of Significance**.
- For the Population 1 Sample, enter 163 as the **Number of Items of Interest** and 227 as the **Sample Size**.
- For the Population 2 Sample, enter 154 as the **Number of Items of Interest** and 262 as the **Sample Size**.
- Click **Two-Tail Test**.
- Enter a **Title** and click **OK**.

In-Depth Excel Use the **COMPUTE worksheet** of the **Z Two Proportions workbook** as a template.

The worksheet already contains data for the hotel guest satisfaction survey. For other problems, change the hypothesized difference, the level of significance, and the number of items of interest and sample size for each group in the cell range B4:B11.

Use the similar **COMPUTE_LOWER** and **COMPUTE_UPPER worksheets** in the same workbook as templates for performing one-tail Z tests for the difference between two proportions. If you use an Excel version older than Excel 2010, use the **COMPUTE_Older** worksheet as a template for both the two-tail and one-tail tests.

Confidence Interval Estimate for the Difference Between Two Proportions

PHStat Modify the **PHStat** instructions for the Z test for the difference between two proportions. In step 6, also check **Confidence Interval Estimate** and enter a **Confidence Level** in its box, in addition to entering a **Title** and clicking **OK**.

In-Depth Excel Use the *In-Depth Excel* instructions for the Z test for the difference between two proportions. The **Z Two Proportions** workbook worksheets include a confidence interval estimate for the difference between two means in the cell range D3:E16.

EG10.4 F TEST for the RATIO of TWO VARIANCES

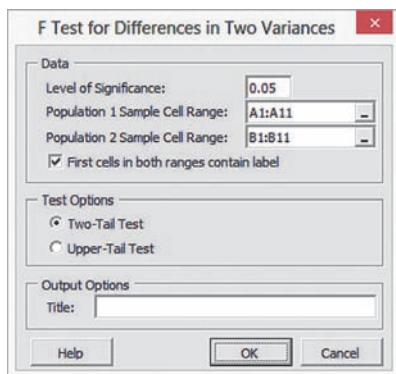
Key Technique Use the **F.INV.RT** (*level of significance / 2, population 1 sample degrees of freedom, population 2 sample degrees of freedom*) function to compute the upper critical value and use the **F.DIST.RT** (*F test statistic, population 1 sample degrees of freedom, population 2 sample degrees of freedom*) function to compute the p -values.

Example Perform the Figure 10.14 F test for the ratio of two variances for the two end-cap locations data shown on page 376.

PHStat Use F Test for Differences in Two Variances.

For the example, open to the **DATA worksheet** of the **Cola** workbook. Select **PHStat** → **Two-Sample Tests (Unsummarized Data)** → **F Test for Differences in Two Variances**. In the procedure's dialog box (shown below):

1. Enter **0.05** as the **Level of Significance**.
2. Enter **A1:A11** as the **Population 1 Sample Cell Range**.
3. Enter **B1:B11** as the **Population 2 Sample Cell Range**.
4. Check **First cells in both ranges contain label**.
5. Click **Two-Tail Test**.
6. Enter a **Title** and click **OK**.



When using summarized data, select **PHStat** → **Two-Sample Tests (Summarized Data)** → **F Test for Differences in Two Variances**. In that procedure's dialog box, enter the level of significance and the sample size and sample variance for each sample.

In-Depth Excel Use the **COMPUTE worksheet** of the **F Two Variances** workbook as a template.

The worksheet already contains the data and formulas for using the unsummarized data for the example. For unsummarized data, paste the data in columns A and B in the **DATACOPY worksheet** and keep the COMPUTE worksheet formulas that compute the sample size and sample variance for the two samples in cell range B4:B10. For summarized data, replace the COMPUTE worksheet formulas in cell ranges B4:B10 with the sample statistics and ignore the DATACOPY worksheet.

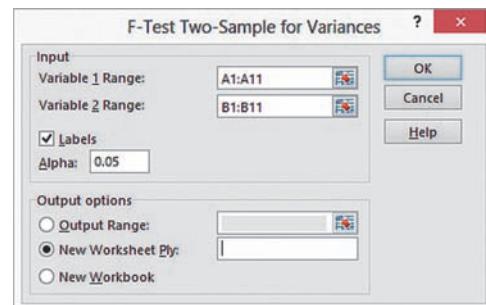
Use the similar **COMPUTE_UPPER worksheet** in the same workbook as a template for performing the upper-tail test. If you

use an Excel version older than Excel 2010, use the **COMPUTE_OLDER** worksheet as a template for both the two-tail and upper-tail tests.

Analysis ToolPak Use F-Test Two-Sample for Variances.

For the example, open to the **DATA worksheet** of the **Cola** workbook and:

1. Select **Data** → **Data Analysis**.
 2. In the Data Analysis dialog box, select **F-Test Two-Sample for Variances** from the **Analysis Tools** list and then click **OK**.
- In the procedure's dialog box (shown below):
3. Enter **A1:A11** as the **Variable 1 Range** and enter **B1:B11** as the **Variable 2 Range**.
 4. Check **Labels** and enter **0.05** as **Alpha**.
 5. Click **New Worksheet Ply**.
 6. Click **OK**.



Results (shown below) appear in a new worksheet and include only the one-tail test *p*-value (0.1241), which must be doubled for the two-tail test shown in Figure 10.14 on page 376.

	A	B	C
1 F-Test Two-Sample for Variances			
2			
3		Beverage	Produce
4 Mean	50.3	72	
5 Variance	350.6778	157.3333	
6 Observations	10	10	
7 df	9	9	
8 F	2.2289		
9 P(F<=f) one-tail	0.1241		
10 F Critical one-tail	3.1789		

CHAPTER 10 MINITAB GUIDE

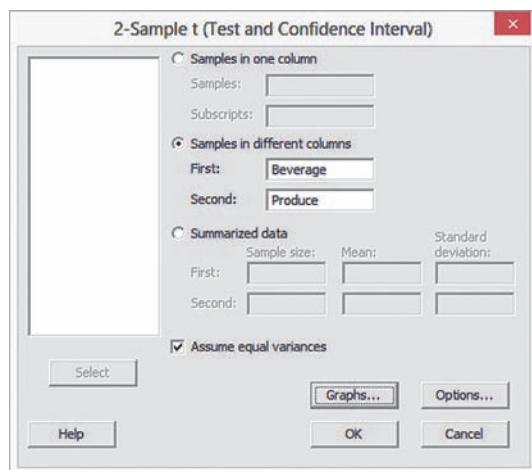
MG10.1 COMPARING the MEANS of TWO INDEPENDENT POPULATIONS

Pooled-Variance t Test for the Difference Between Two Means

Use **2-Sample t**.

For example, to perform the Figure 10.3 pooled-variance *t* test for the two end-cap locations shown on page 350, open to the **Cola worksheet**. Select **Stat → Basic Statistics → 2-Sample t**. In the 2-Sample t (Test and Confidence Interval) dialog box (shown below):

1. Click **Samples in different columns** and press **Tab**.
2. Double-click **C1 Beverage** in the variables list to add **Beverage** to the **First** box.
3. Double-click **C2 Produce** in the variables list to add **Produce** to the **Second** box.
4. Check **Assume equal variances**.
5. Click **Options**.



In the 2-Sample t-Options dialog box (not shown):

6. Enter **95.0** in the **Confidence level** box.
7. Select **not equal** from the **Alternative** drop-down list.
8. Click **OK**.
9. Back in the original dialog box, click **OK**.

For stacked data, use these replacement steps 1 through 3:

1. Click **Samples in one column**.
2. Enter the name of the column that contains the measurement in the **Samples** box.
3. Enter the name of the column that contains the sample names in the **Subscripts** box.

To create a boxplot for the analysis, replace step 9 with the following steps 9 through 11:

9. Back in the original dialog box, click **Graphs**.
10. In the 2-Sample t-Graphs dialog box (not shown), check **Boxplots of data** and then click **OK**.
11. Back in the original dialog box, click **OK**.

For a one-tail test, select **less than** or **greater than** in step 7.

Confidence Interval Estimate for the Difference Between Two Means

Use the instructions for the pooled-variance *t* test, which computes a confidence interval estimate as part of the analysis.

t Test for the Difference Between Two Means, Assuming Unequal Variances

Use the instructions for the pooled-variance *t* test with this replacement step 4:

4. Clear **Assume equal variances**.

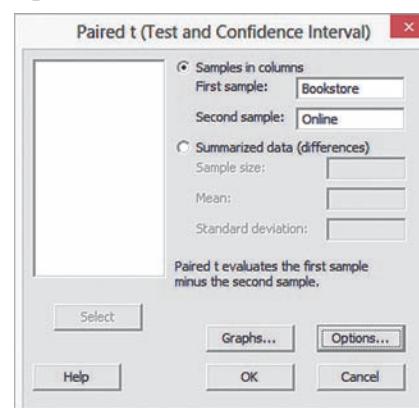
MG10.2 COMPARING the MEANS of TWO RELATED POPULATIONS

Paired t Test

Use **Paired t**.

For example, to perform the Figure 10.9 paired *t* test for the textbook price data on page 363, open to the **BookPrices worksheet**. Select **Stat → Basic Statistics → Paired t**. In the Paired *t* (Test and Confidence Interval) dialog box (shown below):

1. Click **Samples in columns** and press **Tab**.
2. Double-click **C3 Bookstore** in the variables list to enter **Bookstore** in the **First sample** box.
3. Double-click **C4 Online** in the variables list to enter **Online** in the **Second sample** box.
4. Click **Options**.



In the Paired t-Options dialog box (not shown):

5. Enter **95.0** in the **Confidence level** box.
6. Select **not equal** from the **Alternative** drop-down list.
7. Click **OK**.
8. Back in the original dialog box, click **OK**.

To create a boxplot, replace step 8 with the following steps 8 through 10:

8. Back in the original dialog box, click **Graphs**.
9. In the Paired t-Graphs dialog box (not shown), check **Box-plots of data** and then click **OK**.
10. Back in the original dialog box, click **OK**.

For a one-tail test, select **less than** or **greater than** in step 6.

Confidence Interval Estimate for the Mean Difference

Use the instructions for the paired *t* test, which computes a confidence interval estimate as part of the analysis.

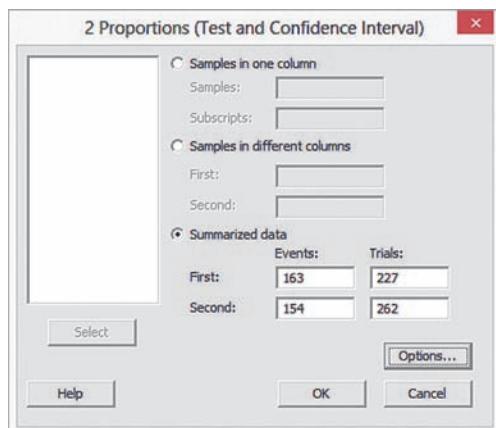
MG10.3 COMPARING the PROPORTIONS of TWO INDEPENDENT POPULATIONS

Z Test for the Difference Between Two Proportions

Use 2 Proportions.

For example, to perform the Figure 10.13 Z test for the hotel guest satisfaction survey on page 370, select **Stat → Basic Statistics → 2 Proportions**. In the 2 Proportions (Test and Confidence Interval) dialog box (shown below):

1. Click **Summarized data**.
2. In the **First** row, enter **163** in the **Events** box and **227** in the **Trials** box.
3. In the **Second** row, enter **154** in the **Events** box and **262** in the **Trials** box.
4. Click **Options**.



In the 2 Proportions - Options dialog box (shown in the right column):

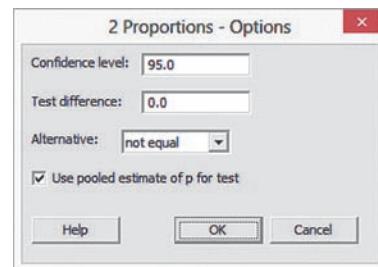
5. Enter **95.0** in the **Confidence level** box.
6. Enter **0.0** in the **Test difference** box.

7. Select **not equal** from the **Alternative** drop-down list.

8. Check **Use pooled estimate of p for test**.

9. Click **OK**.

10. Back in the 2 Proportions (Test and Confidence Interval) dialog box, click **OK**.



Confidence Interval Estimate for the Difference Between Two Proportions

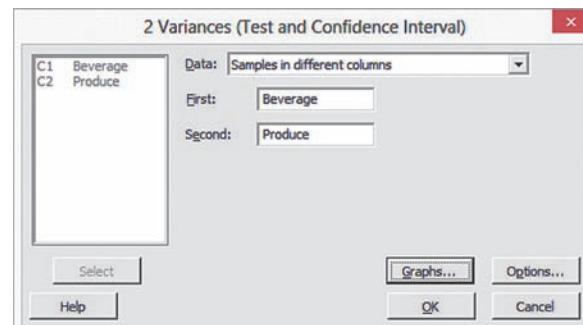
Use the instructions for the Z test for the difference between two proportions, which computes a confidence interval estimate as part of the analysis.

MG10.4 F TEST for the RATIO of TWO VARIANCES

Use 2 Variances.

For example, to perform the Figure 10.14 *F* test for the two end-cap locations on page 376, open to the **COLA worksheet**. Select **Stat → Basic Statistics → 2 Variances**. In the 2 Variances (Test and Confidence Interval) dialog box (shown below):

1. Select **Samples in different columns** from the **Data** drop-down list and press **Tab**.
2. Double-click **C1 Beverage** in the variables list to add **Beverage** to the **First** box.
3. Double-click **C2 Produce** in the variables list to add **Produce** to the **Second** box.
4. Click **Graphs**.



In the 2 Variances - Graphs dialog box (not shown):

5. Clear all check boxes.
6. Click **OK**.
7. Back in the 2 Variances (Test and Confidence Interval) dialog box, click **OK**.

For summarized data, select **Sample standard deviations** or **Sample variances** in step 1 and enter the sample size and the sample statistics for the two variables in lieu of steps 2 and 3.

For stacked data, use these replacement steps 1 through 3:

1. Select **Samples in one column** from the **Data** drop-down list.
2. Enter the name of the column that contains the measurement in the **Samples** box.

3. Enter the name of the column that contains the sample names in the **Subscripts** box.

If you use an older version of Minitab, you will see a 2 Variances dialog box instead of the 2 Variances (Test and Confidence Interval) dialog box. This older dialog box is similar, and you click either **Samples in different columns** or **Summarized data** and then make entries similar to the ones listed in this section. The results created will differ slightly from the results shown in Figure 10.14.

CHAPTER

11

Analysis of Variance

CONTENTS

- 11.1 The Completely Randomized Design: One-Way ANOVA
- 11.2 The Randomized Block Design
- 11.3 The Factorial Design: Two-Way ANOVA
- 11.4 Fixed Effects, Random Effects, and Mixed Effects Models (*online*)

USING STATISTICS: The Means to Find Differences at Arlington's Revisited

CHAPTER 11 EXCEL GUIDE

CHAPTER 11 MINITAB GUIDE

OBJECTIVES

- To introduce the basic concepts of experimental design
- To learn to use the one-way analysis of variance to test for differences among the means of several groups
- To learn when and how to use a randomized block design
- To learn to use the two-way analysis of variance and interpret the interaction effect
- To learn to perform multiple comparisons in a one-way analysis of variance, a randomized block design, and a two-way analysis of variance

USING STATISTICS

The Means to Find Differences at Arlington's

The senior management of Arlington's, a general merchandiser that competes with discount and wholesale club retailers, has just completed a new strategic plan. Among other things, that plan identifies the boosting of mobile electronics sales as an important goal for the future. As a member of the sales team, you are eager to begin implementing this goal, especially as a key regional competitor, Whitney Wireless, has floundered as it awaits integration with the electronics retailer Good Tunes & More (see the Chapter 1 Using Statistics scenario).

Your team has already explored a number of possibilities but has decided to explore how different in-store locations might affect the sales of mobile electronics merchandise. Your team wants to design an experiment in which mobile electronics in selected stores will be sold in one of these three new in-store locations rather than in the current in-aisle location: at the front of the store near weekly specials, in an end-of-aisle (end-cap) special kiosk display, or adjacent to the Expert Counter that is staffed with specially trained salespeople.

You select 20 Arlington's stores that have similar annual sales and divide the stores into four groups of five stores each. To each group, you assign a different in-store sales location for mobile electronics (either the current in-aisle, "front," "kiosk," or "expert" locations). How would you determine if varying the locations had an effect on mobile electronics sales? If you also wanted to explore the effects of permitting mobile payment methods to buy this type of merchandise, could you design an experiment that examined this second factor as it examined the effects of in-store location?



Pavel L Photo and Video/Shutterstock

Comparing possible differences has been the subject of the statistical methods discussed in the previous two chapters. In the one-sample tests of Chapter 9, the comparison is to a standard, such as a certain mean weight for a cereal box being filled by a production line. In Chapter 10, the comparison is between samples taken from two populations. **Analysis of variance**, known by the acronym **ANOVA**, allows statistical comparison among samples taken from many populations.

In ANOVA, the comparison is typically the result of an experiment. For example, the management of a general merchandiser might be brainstorming ways of improving sales of mobile electronics items. At Arlington's, the management decided to try selling those items in four different in-store locations and then observe what the sales would be in each of those locations. The basis for an ANOVA experiment is called the **factor**, which in the Arlington's scenario is in-store location. Used in this way, the statistical use of factor complements everyday business use of the word, such as in the question “How much of a *factor* is in-store location in determining mobile electronics sales?” that the Arlington's sales management team may have asked.

The actual different locations (in-aisle, “front,” “kiosk,” and “expert”) are the **levels** of the factor. Levels of a factor are analogous to the categories of a categorical variable, but you call in-store location a *factor* and not a categorical variable because the variable under study is mobile electronics sales. Levels provide the basis of comparison by dividing the variable under study into **groups**. In the Arlington's scenario, the groups are the stores selling the mobile electronics in-aisle, the stores selling that merchandise at the “front,” the stores selling that merchandise in “kiosks,” and stores selling the merchandise adjacent to the “experts.”

When performing ANOVA analysis, among the types of experiments that can be conducted are:

- Completely randomized design: An experiment with only one factor.
- Randomized block design: An experiment in which the members of each group have been placed in blocks either by being matched or subjected to repeated measurements as was done with the two populations of a paired *t* test (discussed in Section 10.2).
- Factorial design: An experiment in which more than one factor is considered. This chapter discusses two-way ANOVA that involves two factors as an example of this type of design. (Arlington's considering the effects of allowing mobile payments while also experimenting with in-store location would be an example of a factorial design.)

Determining the type of design and the factor or factors the design uses becomes an additional step in the Define task of the DCOVA framework when performing ANOVA analysis.

While ANOVA literally does analyze variation, the purpose of ANOVA is to reach conclusions about possible differences among the *means* of each group, analogous to the hypothesis tests of the previous chapter. Every ANOVA design uses samples that represent each group and subdivides the total variation observed across all samples (all groups) toward the goal of analyzing possible differences among the means of each group. How this subdivision, called *partitioning*, works is a function of the design being used, but total variation, represented by the quantity **sum of squares total (SST)**, will always be the starting point. As with other statistical methods, ANOVA requires making assumptions about the populations that the groups represent. While these assumptions are discussed on page 405 as part of Section 11.1, the assumptions apply for all of the ANOVA methods discussed in this chapter.

Student Tip

ANOVA is also related to regression, a topic discussed later in this book. Because of ANOVA's special relationship with both hypothesis testing and regression, understanding the foundational concepts of ANOVA will prove very helpful in understanding other types of analysis presented in Chapters 13 through 17.

11.1 The Completely Randomized Design: One-Way ANOVA

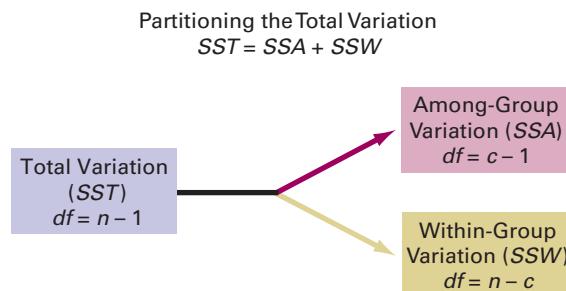
The **completely randomized design** is the ANOVA method that analyzes a single factor. You execute this design using the statistical method **one-way ANOVA**. One-way ANOVA is a two-part process. You first determine if there is a significant difference among the group means. If you reject the null hypothesis that there is no difference among the means, you continue with a second method that seeks to identify the groups whose means are significantly different from the other group means.

Analyzing Variation in One-Way ANOVA

In one-way ANOVA, to analyze variation towards the goal of determining possible differences among the group means, you partition the total variation into variation that is due to differences among the groups and variation that is due to differences within the groups (see Figure 11.1). The **within-group variation (SSW)** measures random variation. The **among-group variation (SSA)** measures differences from group to group. The symbol n represents the number of values in all groups and the symbol c represents the number of groups.

FIGURE 11.1

Partitioning the total variation in a completely randomized design



If using Excel, always organize multiple-sample data as unstacked data, one column per group. (Some Minitab procedures work best with stacked data.) For more information about unstacked (and stacked) data, see page 49.



Student Tip

Another way of stating the alternative hypothesis, H_1 , is that at least one population mean is different from the others.

Assuming that the c groups represent populations whose values are randomly and independently selected, follow a normal distribution, and have equal variances, the null hypothesis of no differences in the population means:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_c$$

is tested against the alternative that not all the c population means are equal:

$$H_1: \text{Not all } \mu_j \text{ are equal (where } j = 1, 2, \dots, c\text{)}.$$

To perform an ANOVA test of equality of population means, you subdivide the total variation in the values into two parts—that which is due to variation among the groups and that which is due to variation within the groups. The **total variation** is represented by the **sum of squares total (SST)**. Because the population means of the c groups are assumed to be equal under the null hypothesis, you compute the total variation among all the values by summing the squared differences between each individual value and the **grand mean**, \bar{X} . The grand mean is the mean of all the values in all the groups combined. Equation (11.1) shows the computation of the total variation.

TOTAL VARIATION IN ONE-WAY ANOVA

$$SST = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 \quad (11.1)$$

where

$$\bar{X} = \frac{\sum_{j=1}^c \sum_{i=1}^{n_j} X_{ij}}{n} = \text{grand mean}$$

X_{ij} = i th value in group j

n_j = number of values in group j

n = total number of values in all groups combined
(that is, $n = n_1 + n_2 + \cdots + n_c$)

c = number of groups

Student Tip

Remember that a sum of squares (SS) cannot be negative.

You compute the among-group variation, usually called the **sum of squares among groups (SSA)**, by summing the squared differences between the sample mean of each group, \bar{X}_j , and the grand mean, \bar{X} , weighted by the sample size, n_j , in each group. Equation (11.2) shows the computation of the among-group variation.

AMONG-GROUP VARIATION IN ONE-WAY ANOVA

$$SSA = \sum_{j=1}^c n_j (\bar{X}_j - \bar{X})^2 \quad (11.2)$$

where

- c = number of groups
- n_j = number of values in group j
- \bar{X}_j = sample mean of group j
- \bar{X} = grand mean

The within-group variation, usually called the **sum of squares within groups (SSW)**, measures the difference between each value and the mean of its own group and sums the squares of these differences over all groups. Equation (11.3) shows the computation of the within-group variation.

WITHIN-GROUP VARIATION IN ONE-WAY ANOVA

$$SSW = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 \quad (11.3)$$

where

- X_{ij} = i th value in group j
- \bar{X}_j = sample mean of group j

Because you are comparing c groups, there are $c - 1$ degrees of freedom associated with the sum of squares among groups. Because each of the c groups contributes $n_j - 1$ degrees of freedom, there are $n - c$ degrees of freedom associated with the sum of squares within groups. In addition, there are $n - 1$ degrees of freedom associated with the sum of squares total because you are comparing each value, X_{ij} , to the grand mean, \bar{X} , based on all n values.

If you divide each of these sums of squares by its respective degrees of freedom, you have three variances, which in ANOVA are known as **mean squares**: MSA (mean square among), MSW (mean square within), and MST (mean square total).

Student Tip

Remember, *mean square* is just another term for *variance* that is used in the analysis of variance. Also, since the mean square is equal to the sum of squares divided by the degrees of freedom, a mean square can never be negative.

MEAN SQUARES IN ONE-WAY ANOVA

$$MSA = \frac{SSA}{c - 1} \quad (11.4a)$$

$$MSW = \frac{SSW}{n - c} \quad (11.4b)$$

$$MST = \frac{SST}{n - 1} \quad (11.4c)$$

F Test for Differences Among More Than Two Means

To determine if there is a significant difference among the group means, you use the *F* test for differences among more than two means. If the null hypothesis is true and there are no differences among the c group means, MSA , MSW , and MST , will provide estimates of the overall variance in the population. Thus, to test the null hypothesis:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_c$$

against the alternative:

$$H_1: \text{Not all } \mu_j \text{ are equal (where } j = 1, 2, \dots, c\text{)}$$

you compute the one-way ANOVA F_{STAT} test statistic as the ratio of MSA to MSW , as in Equation (11.5).

Student Tip

The test statistic compares mean squares (the variances) because one-way ANOVA reaches conclusions about possible differences among the *means* of c groups by examining *variances*.

ONE-WAY ANOVA F_{STAT} TEST STATISTIC

$$F_{STAT} = \frac{MSA}{MSW} \quad (11.5)$$

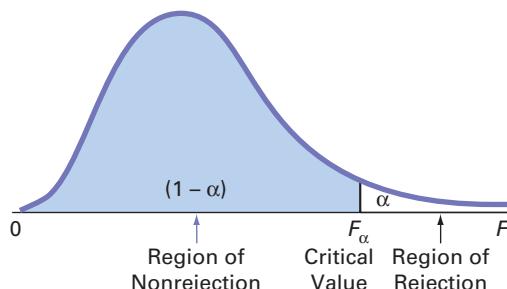
The F_{STAT} test statistic follows an ***F distribution***, with $c - 1$ degrees of freedom in the numerator and $n - c$ degrees of freedom in the denominator.

For a given level of significance, α , you reject the null hypothesis if the F_{STAT} test statistic computed in Equation (11.5) is greater than the upper-tail critical value, F_α , from the *F* distribution with $c - 1$ degrees of freedom in the numerator and $n - c$ in the denominator (see Table E.5). Thus, as shown in Figure 11.2, the decision rule is

Reject H_0 if $F_{STAT} > F_\alpha$;
otherwise, do not reject H_0 .

FIGURE 11.2

Regions of rejection and nonrejection when using ANOVA



If the null hypothesis is true, the computed F_{STAT} test statistic is expected to be approximately equal to 1 because both the numerator and denominator mean square terms are estimating the overall variance in the population. If H_0 is false (and there are differences in the group means), the computed F_{STAT} test statistic is expected to be larger than 1 because the numerator, MSA , is estimating the differences among groups in addition to the overall variability in the values, while the denominator, MSW , is measuring only the overall variability in the values. Therefore, you reject the null hypothesis at a selected level of significance, α , only if the computed F_{STAT} test statistic is *greater than* F_α , the upper-tail critical value of the F distribution having $c - 1$ and $n - c$ degrees of freedom.

Table 11.1 presents the **ANOVA summary table** that is typically used to summarize the results of a one-way ANOVA. The table includes entries for the sources of variation (among groups, within groups, and total), the degrees of freedom, the sums of squares, the mean squares (the variances), and the computed F_{STAT} test statistic. The table may also include the p -value, the probability of having an F_{STAT} value as large as or larger than the one computed, given that the null hypothesis is true. The p -value allows you to reach conclusions about the null hypothesis without needing to refer to a table of critical values of the F distribution. If the p -value is less than the chosen level of significance, α , you reject the null hypothesis.

TABLE 11.1
ANOVA Summary Table

Source	Degrees of Freedom	Sum of Squares	Mean Square (Variance)	F
Among groups	$c - 1$	SSA	$MSA = \frac{SSA}{c - 1}$	$F_{STAT} = \frac{MSA}{MSW}$
Within groups	$n - c$	SSW	$MSW = \frac{SSW}{n - c}$	
Total	$n - 1$	SST		

To illustrate the one-way ANOVA F test, return to the Arlington's scenario (see page 394). You define the business problem as whether significant differences exist in the mobile electronics sales for the four different in-store locations, the four groups for the ANOVA analysis.

To test the comparative effectiveness of the four in-store locations, you conduct a 60-day experiment at 20 same-sized stores that have similar storewide net sales. You randomly assign five stores to use the current in-aisle location, five stores to use the front of the store near weekly specials, five stores to use the end-cap special kiosk display, and five stores to use the location adjacent to the Expert Counter. At the end of the experiment, you organize the mobile electronics sales data by group and store the data in unstacked format in **Mobile Electronics**. Figure 11.3 presents that unstacked data, along with the sample mean and the sample standard deviation for each group.

FIGURE 11.3
Mobile electronic sales (\$000), sample means, and sample standard deviations for four different in-store locations

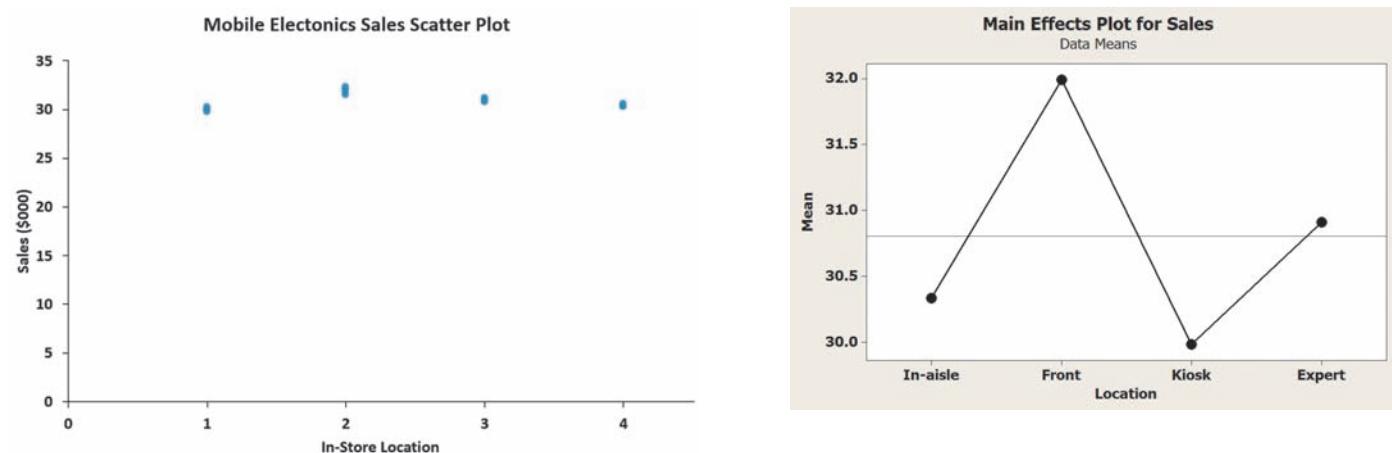
	In-aisle	Front	Kiosk	Expert
1	30.06	32.22	30.78	30.33
2	29.96	31.47	30.91	30.29
3	30.19	32.13	30.79	30.25
4	29.96	31.86	30.95	30.25
5	29.74	32.29	31.13	30.55
Sample Mean	29.982	31.994	30.912	30.334
Sample Standard Deviation	0.165	0.335	0.143	0.125

Figure 11.3 shows differences among the sample means for the mobile electronics sales for the four in-store locations. For the original in-aisle location, mean sales were \$29.982 thousands, whereas mean sales at the three new locations varied from \$30.334 thousands (“expert” location) to \$30.912 thousands (“kiosk” location) to \$31.994 thousands (“front” location).

Differences in the mobile electronic sales for the four in-store locations can also be presented visually. In Figure 11.4, the Minitab cell means plot displays the four sample means and connects the sample means with a straight line. In the same figure, the Excel scatter plot presents the mobile electronics sales at each store in each group, permitting you to observe differences *within* each location as well as among the four locations. (In this example, because the difference within each group is slight, the points for each group overlap and blur together.)

FIGURE 11.4

Excel scatter plot and Minitab main effects plot of mobile electronics sales for four in-store locations



In the Excel scatter plot, the locations in-aisle, front, kiosk, and expert were relabeled 1, 2, 3, and 4 in order to use the scatter plot chart type.

Student Tip

If the sample sizes in each group were larger, you could construct stem-and-leaf displays, boxplots, and normal probability plots as additional ways of visualizing the sales data.

Having observed that the four sample means appear to be different, you use the F test for differences among more than two means to determine if these sample means are sufficiently different to conclude that the *population* means are not all equal. The null hypothesis states that there is no difference in the mean sales among the four in-store locations:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

The alternative hypothesis states that at least one of the in-store location mean sales differs from the other means:

$$H_1: \text{Not all the means are equal.}$$

To construct the ANOVA summary table, you first compute the sample means in each group (see Figure 11.3 on page 399). Then you compute the grand mean by summing all 20 values and dividing by the total number of values:

$$\bar{\bar{X}} = \frac{\sum_{j=1}^c \sum_{i=1}^{n_j} X_{ij}}{n} = \frac{616.12}{20} = 30.806$$

Then, using Equations (11.1) through (11.3) on pages 396–397, you compute the sum of squares:

$$\begin{aligned} SSA &= \sum_{j=1}^c n_j (\bar{X}_j - \bar{\bar{X}})^2 = (5)(29.982 - 30.806)^2 + (5)(31.994 - 30.806)^2 \\ &\quad + (5)(30.912 - 30.806)^2 + (5)(30.334 - 30.806)^2 \\ &= 11.6217 \end{aligned}$$

$$\begin{aligned}
 SSW &= \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 \\
 &= (30.06 - 29.982)^2 + \cdots + (29.74 - 29.982)^2 + (32.22 - 31.994)^2 + \cdots \\
 &\quad + (32.29 - 31.994)^2 + (30.78 - 30.912)^2 + \cdots + (31.13 - 30.912)^2 \\
 &\quad + (30.33 - 30.334)^2 + \cdots + (30.55 - 30.334)^2 \\
 &= 0.7026 \\
 SST &= \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 \\
 &= (30.06 - 30.806)^2 + (29.96 - 30.806)^2 + \cdots + (30.55 - 30.806)^2 \\
 &= 12.3243
 \end{aligned}$$

You compute the mean squares by dividing the sum of squares by the corresponding degrees of freedom [see Equation (11.4) on page 398]. Because $c = 4$ and $n = 20$,

$$\begin{aligned}
 MSA &= \frac{SSA}{c - 1} = \frac{11.6217}{4 - 1} = 3.8739 \\
 MSW &= \frac{SSW}{n - c} = \frac{0.7026}{20 - 4} = 0.0439
 \end{aligned}$$

so that using Equation (11.5) on page 398,

$$F_{STAT} = \frac{MSA}{MSW} = \frac{3.8739}{0.0439} = 88.2186$$

Because you are trying to determine whether MSA is greater than MSW , you only reject H_0 if F_{STAT} is greater than the upper critical value of F . For a selected level of significance, α , you find the upper-tail critical value, F_α , from the F distribution using Table E.5. A portion of Table E.5 is presented in Table 11.2. In the in-store location sales experiment, there are 3 degrees of freedom in the numerator and 16 degrees of freedom in the denominator. F_α , the upper-tail critical value at the 0.05 level of significance, is 3.24.

TABLE 11.2

Finding the Critical Value of F with 3 and 16 Degrees of Freedom at the 0.05 Level of Significance

Denominator df_2	Cumulative Probabilities = 0.95 Upper-Tail Area = 0.05 Numerator df_1								
	1	2	3	4	5	6	7	8	9
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54

Source: Extracted from Table E.5.

Because $F_{STAT} = 88.2186$ is greater than $F_\alpha = 3.24$, you reject the null hypothesis (see Figure 11.5). You conclude that there is a significant difference in the mean sales for the four in-store locations.

FIGURE 11.5

Regions of rejection and nonrejection for the one-way ANOVA at the 0.05 level of significance, with 3 and 16 degrees of freedom

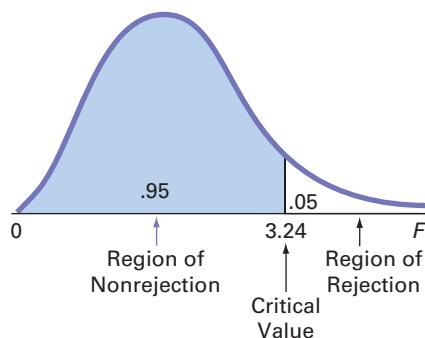


Figure 11.6 shows the ANOVA results for the in-store location sales experiment, including the p -value. In Figure 11.6, what Table 11.1 (see page 399) labels Among Groups is labeled Between Groups in the Excel worksheet. Minitab labels Among Groups as Factor and Within Groups as Error.

FIGURE 11.6

Excel and Minitab ANOVA results for the in-store location sales experiment

A	B	C	D	E	F	G
1 One-Way ANOVA (ANOVA: Single Factor)						
2						
3 SUMMARY						
4 Groups Count Sum Average Variance						
5 In-aisle	5	149.91	29.982	0.02722		
6 Front	5	159.97	31.994	0.11243		
7 Kiosk	5	154.56	30.912	0.02032		
8 Expert	5	151.67	30.334	0.01568		
9						
10						
11 ANOVA						
12 Source of Variation	SS	df	MS	F	P-value	F crit
13 Between Groups	11.6217	3	3.8739	88.2186	0.0000	3.2389
14 Within Groups	0.7026	16	0.0439			
15						
16 Total	12.3243	19				
17					Level of significance	0.05

The formulas in the Excel results worksheet are not shown in Figure 11.6 but are discussed in Section EG11.1 and the SHORT TAKES for Chapter 11.

One-way ANOVA: In-aisle, Front, Kiosk, Expert						
Source	DF	SS	MS	F	P	
Factor	3	11.6217	3.8739	88.22	0.000	
Error	16	0.7026	0.0439			
Total	19	12.3243				
S = 0.2096	R-Sq = 94.30%	R-Sq(adj) = 93.23%				
Individual 95% CIs For Mean Based on Pooled StDev						
Level	N	Mean	StDev	(---*)	(---*)	(---*)
In-aisle	5	29.982	0.165	(---*)		
Front	5	31.994	0.335		(---*)	
Kiosk	5	30.912	0.143			(---*)
Expert	5	30.334	0.125	(---*)		
Pooled StDev = 0.210						

The p -value, or probability of getting a computed F_{STAT} statistic of 88.2186 or larger when the null hypothesis is true, is 0.0000. Because this p -value is less than the specified α of 0.05, you reject the null hypothesis. The p -value of 0.0000 indicates that there is a 0.00% chance of observing differences this large or larger if the population means for the four in-store locations are all equal. After performing the one-way ANOVA and finding a significant difference among the in-store locations, you still do not know *which* in-store locations differ. All you know is that there is sufficient evidence to state that the population means are not all the same. In other words, one or more population means are significantly different. To determine which in-store locations differ, you can use a multiple comparisons procedure such as the Tukey-Kramer procedure.

Student Tip

You have an α level of risk in the entire set of comparisons not just a single comparison.

Multiple Comparisons: The Tukey-Kramer Procedure

In the Arlington's scenario on page 394, you used the one-way ANOVA F test to determine that there was a difference among the suppliers. The next step is to construct **multiple comparisons** to test the null hypothesis that the differences in the means of all pairs of in-store locations are equal to 0.

Although many procedures are available (see references 5, 6, and 10), this text uses the **Tukey-Kramer multiple comparisons procedure for one-way ANOVA** to determine which of the c means are significantly different. This procedure enables you to simultaneously make comparisons between *all* pairs of groups. The procedure consists of the following four steps:

1. Compute the absolute mean differences, $|\bar{X}_j - \bar{X}_{j'}|$ (where j refers to group j , j' refers to group j' , and $j \neq j'$), among all pairs of sample means [$c(c - 1)/2$ pairs].

2. Compute the **critical range** for the Tukey-Kramer procedure, using Equation (11.6). If the sample sizes differ, compute a critical range for each pairwise comparison of sample means.

CRITICAL RANGE FOR THE TUKEY-KRAMER PROCEDURE

$$\text{Critical range} = Q_\alpha \sqrt{\frac{MSW}{2} \left(\frac{1}{n_j} + \frac{1}{n_{j'}} \right)} \quad (11.6)$$

where

n_j = the sample size in group j

$n_{j'}$ = the sample size in group j'

Q_α = the upper-tail critical value from a **Studentized range distribution** having c degrees of freedom in the numerator and $n - c$ degrees of freedom in the denominator.

Student Tip

Table E.7 contains the critical values for the Studentized range distribution.

3. Compare each of the $c(c - 1)/2$ pairs of means against its corresponding critical range. Declare a specific pair significantly different if the absolute difference in the sample means, $|\bar{X}_j - \bar{X}_{j'}|$, is greater than the critical range.
4. Interpret the results.

In the mobile electronics sales example, there are four in-store locations. Thus, there are $4(4 - 1)/2 = 6$ pairwise comparisons. To apply the Tukey-Kramer multiple comparisons procedure, you first compute the absolute mean differences for all six pairwise comparisons:

1. $|\bar{X}_1 - \bar{X}_2| = |29.982 - 31.994| = 2.012$
2. $|\bar{X}_1 - \bar{X}_3| = |29.982 - 30.912| = 0.930$
3. $|\bar{X}_1 - \bar{X}_4| = |29.982 - 30.334| = 0.352$
4. $|\bar{X}_2 - \bar{X}_3| = |31.994 - 30.912| = 1.082$
5. $|\bar{X}_2 - \bar{X}_4| = |31.994 - 30.334| = 1.660$
6. $|\bar{X}_3 - \bar{X}_4| = |30.912 - 30.334| = 0.578$

You then compute only one critical range because the sample sizes in the four groups are equal. (Had the sample sizes in some of the groups been different, you would compute several critical ranges.) From the ANOVA summary table (Figure 11.6 on page 402), $MSW = 0.0439$ and $n_j = n_{j'} = 5$. From Table E.7, for $\alpha = 0.05$, $c = 4$, and $n - c = 20 - 4 = 16$, Q_α , the upper-tail critical value of the test statistic, is 4.05 (see Table 11.3).

TABLE 11.3

Finding the Studentized Range, Q_α , Statistic for $\alpha = 0.05$, with 4 and 16 Degrees of Freedom

Denominator df_2	Cumulative Probabilities = 0.95 Upper-Tail Area = 0.05 Numerator df_1								
	2	3	4	5	6	7	8	9	
:	:	:		:	:	:	:	:	
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	
15	3.01	3.67	4.08	4.37	4.60	4.78	4.94	5.08	
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	

Source: Extracted from Table E.7.

From Equation (11.6),

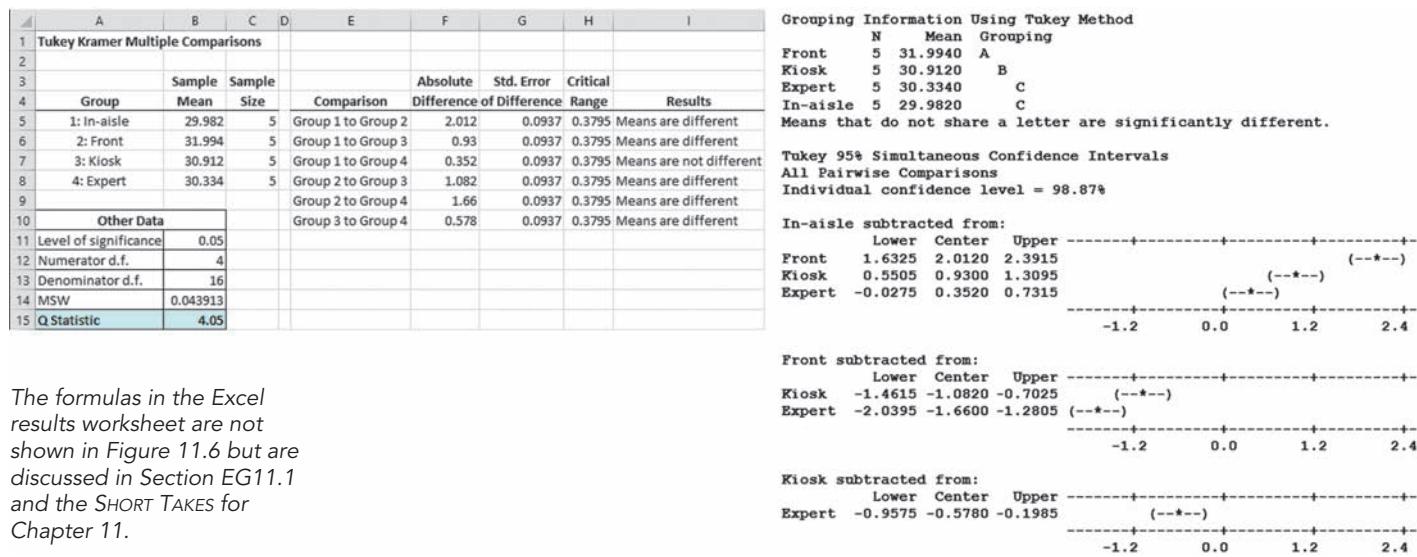
$$\text{Critical range} = 4.05 \sqrt{\left(\frac{0.0439}{2}\right)\left(\frac{1}{5} + \frac{1}{5}\right)} = 0.3795$$

Because the absolute mean difference for five pairs (1, 2, 4, 5, and 6) is greater than 0.3795, you can conclude that there is a significant difference between the mobile electronic sales means of those pairs. Because the absolute mean difference for pair 3 (in-aisle and expert locations) is 0.352, which is less than 0.3795, you conclude that there is no evidence of a difference in the means of those two locations. These results allow you to estimate that the population mean sales for mobile electronics items will be higher at the front location than any other location *and* that the population mean sales for mobile electronics items at kiosk locations will be higher when compared to either the in-aisle or expert locations. As a member of the Arlington's sales team, you would conclude that further study and experimentation with the in-store location of mobile electronics items is appropriate.

Figure 11.7 presents the Excel and Minitab results for the Tukey-Kramer procedure for the mobile electronics sales in-store location experiment. Note that by using $\alpha = 0.05$, you are able to make all six of the comparisons with an overall error rate of only 5%.

FIGURE 11.7

Excel and Minitab Tukey-Kramer procedure results for the in-store location sales experiment



The Figure 11.7 Excel results follow the steps used on pages 402–403 for evaluating the comparisons. The Minitab results show the comparisons in terms of interval estimates. Each interval is computed. Any interval that does not include 0 is considered significant. Thus all the comparisons are significant except for the comparison of in-aisle to expert store location. The interval for that comparison includes 0 since the lower limit is -0.0275 and the upper limit is 0.7315 .

The Analysis of Means (ANOM)

The analysis of means (ANOM) provides an alternative approach that allows you to determine which, if any, of the c groups has a mean significantly different from the overall mean of all the group means combined. The **ANOM online topic** explains this alternative approach and illustrates its use.

Student Tip

To use the one-way ANOVA F test, the variable to be analyzed must either be interval or ratio scaled.

ANOVA Assumptions

In Chapters 9 and 10, you learned about the assumptions required in order to use each hypothesis-testing procedure and the consequences of departures from these assumptions. To use the one-way ANOVA F test, you must make the following assumptions about the populations:

- Randomness and independence
- Normality
- Homogeneity of variance

The first assumption, **randomness and independence**, is critically important. The validity of any experiment depends on random sampling and/or the randomization process. To avoid biases in the outcomes, you need to select random samples from the c groups or use the randomization process to randomly assign the items to the c levels of the factor. Selecting a random sample or randomly assigning the levels ensures that a value from one group is independent of any other value in the experiment. Departures from this assumption can seriously affect inferences from the ANOVA. These problems are discussed more thoroughly in references 5 and 10.

The second assumption, **normality**, states that the sample values in each group are from a normally distributed population. Just as in the case of the t test, the one-way ANOVA F test is fairly robust against departures from the normal distribution. As long as the distributions are not extremely different from a normal distribution, the level of significance of the ANOVA F test is usually not greatly affected, particularly for large samples. You can assess the normality of each of the c samples by constructing a normal probability plot or a boxplot.

The third assumption, **homogeneity of variance**, states that the variances of the c groups are equal (i.e., $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_c^2$). If you have equal sample sizes in each group, inferences based on the F distribution are not seriously affected by unequal variances. However, if you have unequal sample sizes, unequal variances can have a serious effect on inferences from the ANOVA procedure. Thus, when possible, you should have equal sample sizes in all groups. You can use the Levene test for homogeneity of variance to test whether the variances of the c groups are equal.

When only the normality assumption is violated, you can use the Kruskal-Wallis rank test, a nonparametric procedure discussed in Section 12.5. When only the homogeneity-of-variance assumption is violated, you can use procedures similar to those used in the separate-variance t test of Section 10.1 (see references 1 and 2). When both the normality and homogeneity-of-variance assumptions have been violated, you need to use an appropriate data transformation that both normalizes the data and reduces the differences in variances (see reference 6) or use a more general nonparametric procedure (see references 2 and 3).

Levene Test for Homogeneity of Variance

Although the one-way ANOVA F test is relatively robust with respect to the assumption of equal group variances, large differences in the group variances can seriously affect the level of significance and the power of the F test. One powerful yet simple procedure for testing the equality of the variances is the modified **Levene test** (see references 1 and 7). To test for the homogeneity of variance, you use the following null hypothesis:

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_c^2$$

against the alternative hypothesis:

$$H_1: \text{Not all } \sigma_j^2 \text{ are equal } (j = 1, 2, 3, \dots, c)$$

To test the null hypothesis of equal variances, you first compute the absolute value of the difference between each value and the median of the group. Then you perform a one-way ANOVA on these *absolute differences*. Most statisticians suggest using a level of significance of $\alpha = 0.05$ when performing the ANOVA. To illustrate the modified Levene test, return to the Figure 11.3 data and summary statistics on page 399 for the Arlington's scenario concerning

Student Tip

Remember when performing the Levene test that you are conducting a one-way ANOVA on the absolute differences from the median in each group, not on the actual values themselves.

the in-store location sales experiment. Table 11.4 summarizes the absolute differences from the median of each location.

TABLE 11.4

Absolute Differences from the Median Sales for Four Locations

In-Aisle (Median = 29.96)	Front (Median = 32.13)	Kiosk (Median = 30.91)	Expert (Median = 30.29)
$ 30.06 - 29.96 = 0.10$	$ 32.22 - 32.13 = 0.09$	$ 30.78 - 30.91 = 0.13$	$ 30.33 - 30.29 = 0.04$
$ 29.96 - 29.96 = 0.00$	$ 31.47 - 32.13 = 0.66$	$ 30.91 - 30.91 = 0.00$	$ 30.29 - 30.29 = 0.00$
$ 30.19 - 29.96 = 0.23$	$ 32.13 - 32.13 = 0.00$	$ 30.79 - 30.91 = 0.12$	$ 30.25 - 30.29 = 0.04$
$ 29.96 - 29.96 = 0.00$	$ 31.86 - 32.13 = 0.27$	$ 30.95 - 30.91 = 0.04$	$ 30.25 - 30.29 = 0.04$
$ 29.74 - 29.96 = 0.22$	$ 32.29 - 32.13 = 0.16$	$ 31.13 - 30.91 = 0.22$	$ 30.55 - 30.29 = 0.26$

Using the absolute differences given in Table 11.4, you perform a one-way ANOVA (see Figure 11.8).

FIGURE 11.8

Excel and Minitab Levene test results for the absolute differences for the in-store location sales experiment

A	B	C	D	E	F	G
1 ANOVA: Levene Test						
2						
3 SUMMARY						
4 Groups	Count	Sum	Average	Variance		
5 In-aisle	5	0.55	0.11	0.0127		
6 Front	5	1.18	0.236	0.06593		
7 Kiosk	5	0.51	0.102	0.00732		
8 Expert	5	0.38	0.076	0.01088		
9						
10						
11 ANOVA						
12 Source of Variation	SS	df	MS	F	P-value	F crit
13 Between Groups	0.07666	3	0.0256	1.0556	0.3953	3.2389
14 Within Groups	0.38732	16	0.0242			
15						
16 Total	0.46398	19				
17				Level of significance	0.05	

Test for Equal Variances: Sales versus Location
Levene's Test (Any Continuous Distribution)
Test statistic = 1.06, p-value = 0.395

From the Figure 11.8 results, observe that $F_{STAT} = 1.0556$. (The Excel worksheet labels this value F and Minitab labels the value Test statistic.) Because $F_{STAT} = 1.0556 < 3.2389$ (or the p -value = 0.3953 > 0.05), you do not reject H_0 . There is insufficient evidence of a significant difference among the four variances. In other words, it is reasonable to assume that the four in-store locations have an equal amount of variability in sales. Therefore, the homogeneity-of-variance assumption for the ANOVA procedure is justified.

Example 11.1 illustrates another example of the one-way ANOVA.

EXAMPLE 11.1

ANOVA of the Speed of Drive-Through Service at Fast-Food Chains

For fast-food restaurants, the drive-through window is an important revenue source. The chain that offers the fastest service is likely to attract additional customers. Each year *QSR Magazine*, www.qsrmagazine.com, publishes its results of a survey of drive-through service times (from menu board to departure) at fast-food chains. In a recent year, the mean time was 129.75 seconds for Wendy's, 149.69 seconds for Taco Bell, 201.33 seconds for Burger King, 188.83 seconds for McDonald's, and 190.06 seconds for Chick-fil-A. Suppose the study was based on 20 customers for each fast-food chain. At the 0.05 level of significance, is there evidence of a difference in the mean drive-through service times of the five chains?

Table 11.5 contains the ANOVA table for this problem.

TABLE 11.5

ANOVA Summary Table of Drive-Through Service Times at Fast-Food Chains

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F	p-value
Among chains	4	75,048.74	18,762.185	143.66	0.0000
Within chains	95	12,407.00	130.60		

SOLUTION

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ where 1 = Wendy's, 2 = Taco Bell, 3 = Burger King, 4 = McDonald's, 5 = Chick-fil-A

$H_1:$ Not all μ_j are equal where $j = 1, 2, 3, 4, 5$

Decision rule: If the p -value < 0.05 , reject H_0 . Because the p -value is 0.0000, which is less than $\alpha = 0.05$, reject H_0 . You have sufficient evidence to conclude that the mean drive-through times of the five chains are not all equal.

To determine which of the means are significantly different from one another, use the Tukey-Kramer procedure [Equation (11.6) on page 403] to establish the critical range:

Critical value of Q with 5 and 95 degrees of freedom ≈ 3.92

$$\begin{aligned} \text{Critical range} &= Q_\alpha \sqrt{\left(\frac{MSW}{2}\right)\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)} = (3.92) \sqrt{\left(\frac{130.6}{2}\right)\left(\frac{1}{20} + \frac{1}{20}\right)} \\ &= 10.02 \end{aligned}$$

Any observed difference greater than 10.02 is considered significant. The mean drive-through service times are different between Wendy's (mean of 129.75 seconds) and Taco Bell, Burger King, McDonald's, and Chick-fil-A and also between Taco Bell (mean of 149.69) and Burger King, McDonald's, and Chick-fil-A. In addition, the mean drive-through service time is different between Burger King and McDonald's, and between Burger King and Chick-fil-A. Thus, with 95% confidence, you can conclude that the estimated population mean drive-through service time is faster for Wendy's than for Taco Bell. In addition, the population mean service time for Wendy's and for Taco Bell is faster than those of Burger King, McDonald's, and Chick-fil-A. Also, the population mean drive-through service time for Burger King is slower than for McDonald's and for Chick-fil-A.

Problems for Section 11.1

LEARNING THE BASICS

11.1 An experiment has a single factor with five groups and seven values in each group.

- a. How many degrees of freedom are there in determining the among-group variation?
- b. How many degrees of freedom are there in determining the within-group variation?
- c. How many degrees of freedom are there in determining the total variation?

11.2 You are working with the same experiment as in Problem 11.1.

- a. If $SSA = 60$ and $SST = 210$, what is SSW ?
- b. What is MSA ?
- c. What is MSW ?
- d. What is the value of F_{STAT} ?

11.3 You are working with the same experiment as in Problems 11.1 and 11.2.

- a. Construct the ANOVA summary table and fill in all values in the table.
- b. At the 0.05 level of significance, what is the upper-tail critical value from the F distribution?
- c. State the decision rule for testing the null hypothesis that all five groups have equal population means.
- d. What is your statistical decision?

11.4 Consider an experiment with three groups, with seven values in each.

- a. How many degrees of freedom are there in determining the among-group variation?
- b. How many degrees of freedom are there in determining the within-group variation?
- c. How many degrees of freedom are there in determining the total variation?

11.5 Consider an experiment with four groups, with eight values in each. For the ANOVA summary table below, fill in all the missing results:

Source	Degrees of Freedom	Sum of Squares	Mean Square (Variance)	F
Among groups	$c - 1 = ?$	$SSA = ?$	$MSA = 80$	$F_{STAT} = ?$
Within groups	$n - c = ?$	$SSW = 560$	$MSW = ?$	
Total	$n - 1 = ?$	$SST = ?$		

11.6 You are working with the same experiment as in Problem 11.5.

- At the 0.05 level of significance, state the decision rule for testing the null hypothesis that all four groups have equal population means.
- What is your statistical decision?
- At the 0.05 level of significance, what is the upper-tail critical value from the Studentized range distribution?
- To perform the Tukey-Kramer procedure, what is the critical range?

APPLYING THE CONCEPTS

11.7 *Accounting Today* identified the top accounting firms in 10 geographic regions across the United States. Even though all 10 regions reported growth in 2012, the Capital, Great Lakes, Mid-Atlantic, and New England regions reported relatively similar combined growths, of 7.2%, 7.72%, 5.00%, and 5.03%, respectively. A characteristic description of the accounting firms in the Capital, Great Lakes, Mid-Atlantic, and New England regions included the number of partners in the firm.

The file **AccountingPartners4** contains the number of partners. (Data extracted from bit.ly/1XZNPr.)

- At the 0.05 level of significance, is there evidence of a difference among the Capital, Great Lakes, Mid-Atlantic, and New England region accounting firms with respect to the mean number of partners?
- If the results in (a) indicate that it is appropriate to do so, use the Tukey-Kramer procedure to determine which regions differ in the mean number of partners. Discuss your findings.

SELF TEST **11.8** The more costly and time consuming it is to export and import, the more difficult it is for local companies to be competitive and to reach international markets. As part of an initial investigation exploring foreign market entry, 10 countries were selected from each of four global regions. The cost associated with importing a standardized cargo of goods by sea transport in these countries (in US\$ per container) is stored in **ForeignMarket2**. (Data extracted from doingbusiness.org/data.)

- At the 0.05 level of significance, is there evidence of a difference in the mean cost of importing across the four global regions?
- If appropriate, determine which global regions differ in mean cost of importing.
- At the 0.05 level of significance, is there evidence of a difference in the variation in cost of importing among the four global regions?
- Which global region(s) should you consider for foreign market entry? Explain.

11.9 A hospital conducted a study of the waiting time in its emergency room. The hospital has a main campus and three satellite locations. Management had a business objective of reducing waiting time for emergency room cases that did not require immediate attention. To study this, a random sample of 15 emergency room cases that did not require immediate attention at each location were selected on a particular day, and the waiting times (measured from check-in to when the patient was called into the clinic area) were collected and stored in **ERWaiting**.

- At the 0.05 level of significance, is there evidence of a difference in the mean waiting times in the four locations?
- If appropriate, determine which locations differ in mean waiting time.

- At the 0.05 level of significance, is there evidence of a difference in the variation in waiting time among the four locations?

11.10 A manufacturer of pens has hired an advertising agency to develop an advertising campaign for the upcoming holiday season. To prepare for this project, the research director decides to initiate a study of the effect of advertising on product perception. An experiment is designed to compare five different advertisements. Advertisement *A* greatly undersells the pen's characteristics. Advertisement *B* slightly undersells the pen's characteristics. Advertisement *C* slightly oversells the pen's characteristics. Advertisement *D* greatly oversells the pen's characteristics. Advertisement *E* attempts to correctly state the pen's characteristics. A sample of 30 adult respondents, taken from a larger focus group, is randomly assigned to the five advertisements (so that there are 6 respondents to each advertisement). After reading the advertisement and developing a sense of "product expectation," all respondents unknowingly receive the same pen to evaluate. The respondents are permitted to test the pen and the plausibility of the advertising copy. The respondents are then asked to rate the pen from 1 to 7 (lowest to highest) on the product characteristic scales of appearance, durability, and writing performance. The *combined* scores of three ratings (appearance, durability, and writing performance) for the 30 respondents, stored in **Pen**, are as follows:

A	B	C	D	E
15	16	8	5	12
18	17	7	6	19
17	21	10	13	18
19	16	15	11	12
19	19	14	9	17
20	17	14	10	14

- At the 0.05 level of significance, is there evidence of a difference in the mean rating of the pens following exposure to five advertisements?
- If appropriate, determine which advertisements differ in mean ratings.
- At the 0.05 level of significance, is there evidence of a difference in the variation in ratings among the five advertisements?
- Which advertisement(s) should you use, and which advertisement(s) should you avoid? Explain.

11.11 *QSR* has been reporting on the largest quick-serve and fast-casual brands in the United States for nearly 15 years. The file **QSR** contains the food segment (burger, chicken, pizza, or sandwich) and U.S. mean sales per unit (\$ thousands) for each of 38 quick-service brands. (Data extracted from bit.ly/16GJIE.)

- At the 0.05 level of significance, is there evidence of a difference in the mean U.S. mean sales per unit (\$ thousands) among the food segments?
- At the 0.05 level of significance, is there a difference in the variation in U.S. average sales per unit (\$ thousands) among the food segments?
- What effect does your result in (b) have on the validity of the results in (a)?

11.12 Researchers conducted a study to determine whether graduates with an academic background in the discipline of leadership studies were better equipped with essential soft skills required to be successful in contemporary organizations than students with no leadership education and/or students with a certificate in leadership. The Teams Skills Questionnaire was used to capture students' self-reported ratings of their soft skills. The researchers found the following:

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
Among groups	2	1.879		
Within groups	297	31.865		
Total	299	33.744		

Group	N	Mean
No coursework in leadership	109	3.290
Certificate in leadership	90	3.362
Degree in leadership	102	3.471

Source: Data Extracted from C. Brungardt, "The Intersection Between Soft Skill Development and Leadership Education," *Journal of Leadership Education*, 10 (Winter 2011): 1–22.

- Complete the ANOVA summary table.
- At the 0.05 level of significance, is there evidence of a difference in the mean soft-skill score reported by different groups?
- If the results in (b) indicate that it is appropriate, use the Tukey-Kramer procedure to determine which groups differ in mean soft-skill score. Discuss your findings.

11.13 A pet food company has a business objective of expanding its product line beyond its current kidney- and shrimp-based cat foods. The company developed two new products, one based on chicken liver and the other based on salmon. The company conducted an experiment to compare the two new products with its two existing ones, as well as a generic beef-based product sold at a supermarket chain.

For the experiment, a sample of 50 cats from the population at a local animal shelter was selected. Ten cats were randomly assigned to each of the five products being tested. Each of the cats was then presented with 3 ounces of the selected food in a dish at feeding time. The researchers defined the variable to be measured as the number of ounces of food that the cat consumed within a 10-minute time interval that began when the filled dish was presented. The results for this experiment are summarized in the table at top right and stored in **CatFood**.

- At the 0.05 level of significance, is there evidence of a difference in the mean amount of food eaten among the various products?
- If appropriate, determine which products appear to differ significantly in the mean amount of food eaten.
- At the 0.05 level of significance, is there evidence of a difference in the variation in the amount of food eaten among the various products?
- What should the pet food company conclude? Fully describe the pet food company's options with respect to the products.

Kidney	Shrimp	Chicken Liver	Salmon	Beef
2.37	2.26	2.29	1.79	2.09
2.62	2.69	2.23	2.33	1.87
2.31	2.25	2.41	1.96	1.67
2.47	2.45	2.68	2.05	1.64
2.59	2.34	2.25	2.26	2.16
2.62	2.37	2.17	2.24	1.75
2.34	2.22	2.37	1.96	1.18
2.47	2.56	2.26	1.58	1.92
2.45	2.36	2.45	2.18	1.32
2.32	2.59	2.57	1.93	1.94

11.14 A sporting goods manufacturing company wanted to compare the distance traveled by golf balls produced using four different designs. Ten balls were manufactured with each design and were brought to the local golf course for the club professional to test. The order in which the balls were hit with the same club from the first tee was randomized so that the pro did not know which type of ball was being hit. All 40 balls were hit in a short period of time, during which the environmental conditions were essentially the same. The results (distance traveled in yards) for the four designs are stored in **Golfball** and shown in the following table:

Design 1	Design 2	Design 3	Design 4
206.32	217.08	226.77	230.55
207.94	221.43	224.79	227.95
206.19	218.04	229.75	231.84
204.45	224.13	228.51	224.87
209.65	211.82	221.44	229.49
203.81	213.90	223.85	231.10
206.75	221.28	223.97	221.53
205.68	229.43	234.30	235.45
204.49	213.54	219.50	228.35
210.86	214.51	233.00	225.09

- At the 0.05 level of significance, is there evidence of a difference in the mean distances traveled by the golf balls with different designs?
- If the results in (a) indicate that it is appropriate to do so, use the Tukey-Kramer procedure to determine which designs differ in mean distances.
- What assumptions are necessary in (a)?
- At the 0.05 level of significance, is there evidence of a difference in the variation of the distances traveled by the golf balls with different designs?
- What golf ball design should the manufacturing manager choose? Explain.

11.2 The Randomized Block Design

Section 11.1 discussed how to use the one-way ANOVA F test to evaluate differences among the means of more than two independent groups. The **randomized block design** evaluates differences among more than two groups in which the members of each group have been placed in blocks either by being matched or subjected to repeated measurements in the same way as in the paired t test (discussed in Section 10.2). Blocking removes variability due to individual differences so that the differences among the groups are more evident.

Although blocks are used in a randomized block design, the focus of the analysis is on the differences among the different groups. As is the case in completely randomized designs, groups are often different levels pertaining to a factor of interest. A randomized block design is often more statistically powerful than a completely randomized design (see references 5, 6, and 10). For example, if the factor of interest is advertising medium, three groups could be the following different levels: television, radio, and newspaper. Using different cities as blocks removes the variability of the different cities from the random error so as to better detect differences among the three advertising mediums.

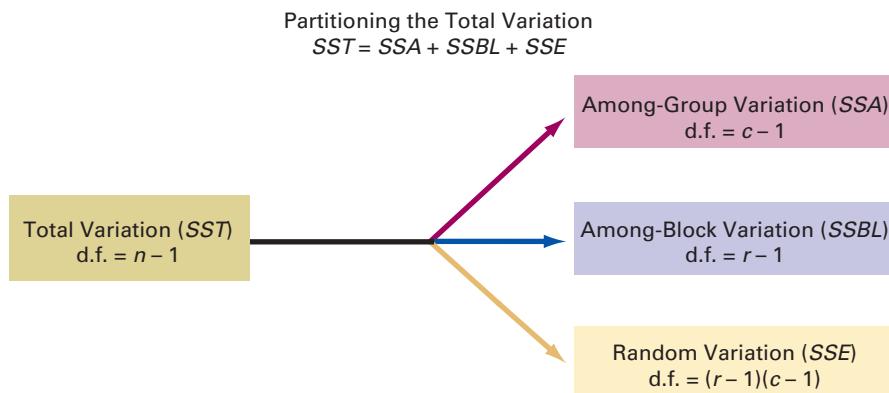
Testing for Factor and Block Effects

Recall from Figure 11.1 on page 396 that, in the completely randomized design, the total variation (SST) is subdivided into variation due to differences *among* the c groups (SSA) and variation due to variation *within* the c groups (SSW). Within-group variation is considered random variation, and among-group variation is due to differences from group to group.

To remove the effects of the blocking from the random variation component in the randomized block design, the within-group variation (SSW) is subdivided into variation due to differences among the blocks ($SSBL$) and random variation (SSE). Therefore, as presented in Figure 11.9, in a randomized block design, the total variation is the sum of three components: among-group variation (SSA), among-block variation ($SSBL$), and random variation (SSE).

FIGURE 11.9

Partitioning the total variation in a randomized block model



The following definitions are needed to develop the ANOVA procedure for the randomized block design:

r = the number of blocks

c = the number of groups

n = the total number of values (where $n = rc$)

X_{ij} = the value in the i th block for the j th group

\bar{X}_i = the mean of all the values in block i

\bar{X}_j = the mean of all the values for group j

$$\sum_{j=1}^c \sum_{i=1}^r X_{ij} = \text{the grand total}$$

The total variation, also called sum of squares total (*SST*), is a measure of the variation among all the values. You compute *SST* by summing the squared differences between each individual value and the grand mean, $\bar{\bar{X}}$, that is based on all n values. Equation (11.7) shows the computation for total variation.

TOTAL VARIATION IN THE RANDOMIZED BLOCK DESIGN

$$SST = \sum_{j=1}^c \sum_{i=1}^r (X_{ij} - \bar{\bar{X}})^2 \quad (11.7)$$

where

$$\bar{\bar{X}} = \frac{\sum_{j=1}^c \sum_{i=1}^r X_{ij}}{rc} \text{ (i.e., the grand mean)}$$

You compute the among-group variation, also called the sum of squares among groups (*SSA*), by summing the squared differences between the sample mean of each group, \bar{X}_j , and the grand mean, $\bar{\bar{X}}$, weighted by the number of blocks, r . Equation (11.8) shows the computation for the among-group variation.

AMONG-GROUP VARIATION IN THE RANDOMIZED BLOCK DESIGN

$$SSA = r \sum_{j=1}^c (\bar{X}_j - \bar{\bar{X}})^2 \quad (11.8)$$

where

$$\bar{X}_j = \frac{\sum_{i=1}^r X_{ij}}{r}$$

You compute the **among-block variation**, also called the **sum of squares among blocks** (*SSBL*), by summing the squared differences between the mean of each block, \bar{X}_i , and the grand mean, $\bar{\bar{X}}$, weighted by the number of groups, c . Equation (11.9) shows the computation for the among-block variation.

AMONG-BLOCK VARIATION IN THE RANDOMIZED BLOCK DESIGN

$$SSBL = c \sum_{i=1}^r (\bar{X}_i - \bar{\bar{X}})^2 \quad (11.9)$$

where

$$\bar{X}_i = \frac{\sum_{j=1}^c X_{ij}}{c}$$

You compute the random variation, also called the **sum of squares error** (*SSE*), by summing the squared differences among all the values after the effect of the groups and blocks have been accounted for. Equation (11.10) shows the computation for random variation.

RANDOM VARIATION IN THE RANDOMIZED BLOCK DESIGN

$$SSE = \sum_{j=1}^c \sum_{i=1}^r (X_{ij} - \bar{X}_j - \bar{X}_i + \bar{\bar{X}})^2 \quad (11.10)$$

Because you are comparing c groups, there are $c - 1$ degrees of freedom associated with the sum of squares among groups (SSA). Similarly, because there are r blocks, there are $r - 1$ degrees of freedom associated with the sum of squares among blocks ($SSBL$). Moreover, there are $n - 1$ degrees of freedom associated with the sum of squares total (SST) because you are comparing each value, X_{ij} , to the grand mean, $\bar{\bar{X}}$, based on all n values. Therefore, because the degrees of freedom for each of the sources of variation must add to the degrees of freedom for the total variation, you compute the degrees of freedom for the sum of squares error (SSE) component by subtraction and algebraic manipulation. Thus, the degrees of freedom associated with the sum of squares error is $(r - 1)(c - 1)$.

If you divide each of the component sums of squares by its associated degrees of freedom, you have the three *variances*, or mean square terms (MSA , $MSBL$, and MSE). Equations (11.11a–c) give the mean square terms needed for the ANOVA table.

THE MEAN SQUARES IN THE RANDOMIZED BLOCK DESIGN

$$MSA = \frac{SSA}{c - 1} \quad (11.11a)$$

$$MSBL = \frac{SSBL}{r - 1} \quad (11.11b)$$

$$MSE = \frac{SSE}{(r - 1)(c - 1)} \quad (11.11c)$$

The first step in analyzing a randomized block design is to test for a factor effect—that is, to test for any differences among the c group means. If the assumptions of the analysis of variance are valid, the null hypothesis of no differences in the c group means:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_c$$

is tested against the alternative that not all the c group means are equal:

$$H_1: \text{Not all } \mu_j \text{ are equal (where } j = 1, 2, \dots, c\text{)}$$

by computing the F_{STAT} test statistic given in Equation (11.12).

F_{STAT} STATISTIC FOR FACTOR EFFECT

$$F_{STAT} = \frac{MSA}{MSE} \quad (11.12)$$

The F_{STAT} test statistic follows an F distribution with $c - 1$ degrees of freedom for the MSA term and $(r - 1)(c - 1)$ degrees of freedom for the MSE term. For a given level of significance α , you reject the null hypothesis if the computed F_{STAT} test statistic is greater than the upper-tail critical value, F_α , from the F distribution with $c - 1$ and $(r - 1)(c - 1)$ degrees of freedom (see Table E.5). The decision rule is:

Reject H_0 if $F_{STAT} > F_\alpha$;

otherwise, do not reject H_0 .

To examine whether the randomized block design was advantageous to use, some statisticians suggest that you perform the F test for block effects. The null hypothesis of no block effects:

$$H_0: \mu_{1.} = \mu_{2.} = \cdots = \mu_{r.}$$

is tested against the alternative:

$$H_1: \text{Not all } \mu_i \text{ are equal (where } i = 1, 2, \dots, r\text{)}$$

using the F_{STAT} test statistic for block effect given in Equation (11.13).

F_{STAT} STATISTIC FOR BLOCK EFFECTS

$$F_{STAT} = \frac{MSBL}{MSE} \quad (11.13)$$

You reject the null hypothesis at the α level of significance if the computed F_{STAT} test statistic is greater than the upper-tail critical value F_α from the F distribution with $r - 1$ and $(r - 1)(c - 1)$ degrees of freedom (see Table E.5). That is, the decision rule is

Reject H_0 if $F_{STAT} > F_\alpha$;

otherwise, do not reject H_0 .

The results of the analysis-of-variance procedure are usually displayed in an ANOVA summary table, as shown in Table 11.6.

TABLE 11.6

Analysis-of-Variance Table for the Randomized Block Design

Source	Degrees of Freedom	Sum of Squares	Mean Square (Variance)	F
Among groups (A)	$c - 1$	SSA	$MSA = \frac{SSA}{c - 1}$	$F_{STAT} = \frac{MSA}{MSE}$
Among blocks (BL)	$r - 1$	$SSBL$	$MSBL = \frac{SSBL}{r - 1}$	$F_{STAT} = \frac{MSBL}{MSE}$
Error Total	$\frac{(r - 1)(c - 1)}{rc - 1}$	$\frac{SSE}{SST}$	$MSE = \frac{SSE}{(r - 1)(c - 1)}$	

To illustrate the randomized block design, suppose that a quick-service chain wants to evaluate the service at four of its restaurants. The customer service director for the chain hires six evaluators with varied experiences in food-service evaluations to act as raters. To reduce the effect of the variability between evaluators, you use a randomized block design, with evaluators serving as the blocks. The four restaurants are the groups of interest.

The six evaluators rate the service at each of the four restaurants in a random order. A rating scale from 0 (low) to 100 (high) is used. Table 11.7 summarizes the results (stored in **QSRChain**), along with the group totals, group means, block totals, block means, grand total, and grand mean.

TABLE 11.7

Service Ratings for Four Restaurants of a Quick-Service Chain

EVALUATORS	RESTAURANT				Totals	Means
	Henry St	Surf Ave	Granby	Blvd N		
1	70	61	82	74	287	71.75
2	77	75	88	76	316	79.00
3	76	67	90	80	313	78.25
4	80	63	96	76	315	78.75
5	84	66	92	84	326	81.50
6	78	68	98	86	330	82.50
Totals	465	400	546	476	1,887	
Means	77.50	66.67	91.00	79.33	78.625	

In addition, from Table 11.7,

$$r = 6 \quad c = 4 \quad n = rc = 24$$

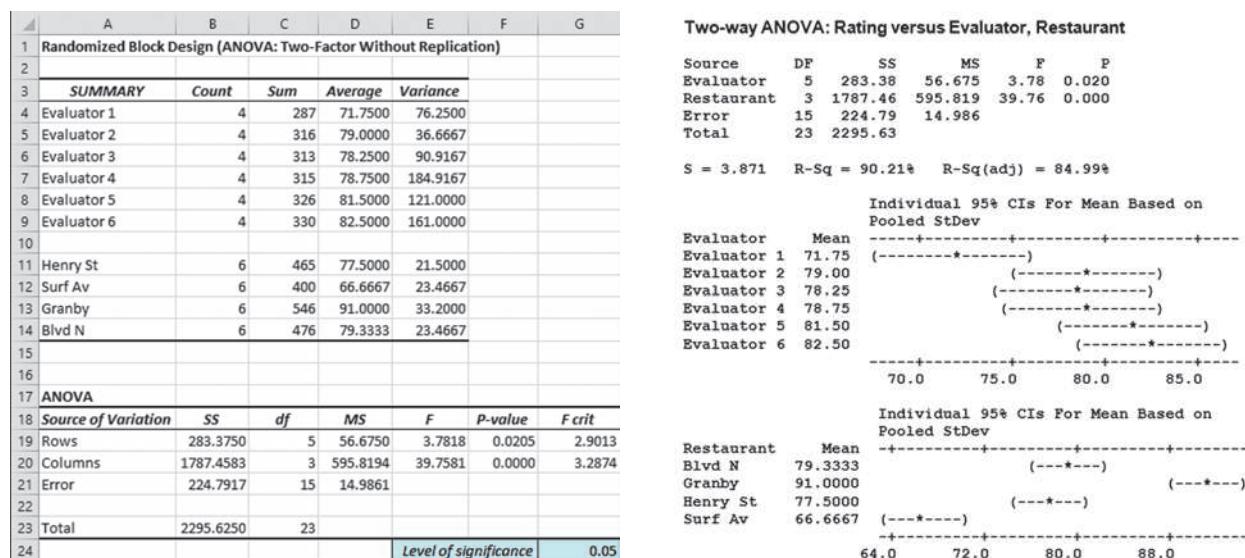
and

$$\bar{\bar{X}} = \frac{\sum_{j=1}^c \sum_{i=1}^r X_{ij}}{rc} = \frac{1,887}{24} = 78.625$$

Figure 11.10 shows the results for this randomized block design. (In the Excel ANOVA table, Rows are the evaluators and Columns are the restaurants.)

FIGURE 11.10

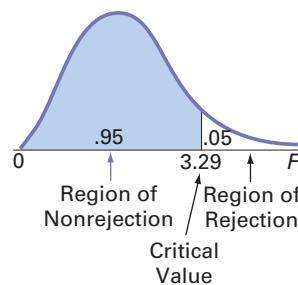
Excel and Minitab randomized block design results for the quick-service chain study



Using the 0.05 level of significance to test for differences among the restaurants, you reject the null hypothesis ($H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$) if the computed F_{STAT} test statistic is greater than 3.29, the upper-tail critical value from the F distribution with 3 and 15 degrees of freedom in the numerator and denominator, respectively (see Figure 11.11).

FIGURE 11.11

Regions of rejection and nonrejection for the quick service chain study at the 0.05 level of significance with 3 and 15 degrees of freedom



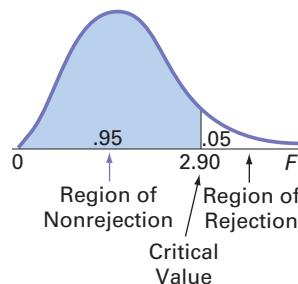
Because $F_{STAT} = 39.7581 > F_\alpha = 3.29$, or because the $p\text{-value} = 0.0000 < 0.05$, you reject H_0 and conclude that there is evidence of a difference in the mean ratings among the different restaurants. The extremely small $p\text{-value}$ indicates that if the means from the four restaurants are equal, the probability is 0.0000 that you will get differences as large or larger among the sample means, as observed in this study. Thus, there is little degree of belief in

the null hypothesis. You conclude that the alternative hypothesis is correct: The mean ratings among the four restaurants are different.

As a check on the effectiveness of blocking, you can test for a difference among the evaluators. The decision rule, using the 0.05 level of significance, is to reject the null hypothesis ($H_0: \mu_1 = \mu_2 = \dots = \mu_6$) if the computed F_{STAT} test statistic is greater than 2.90, the upper-tail critical value from the F distribution with 5 and 15 degrees of freedom (see Figure 11.12). Because $F_{STAT} = 3.7818 > F_\alpha = 2.90$ or because the p -value = 0.0205 < 0.05, you reject H_0 and conclude that there is evidence of a difference among the evaluators. Thus, you conclude that the blocking has been advantageous in reducing the random error.

FIGURE 11.12

Regions of rejection and nonrejection for the quick-service chain study at the 0.05 level of significance with 5 and 15 degrees of freedom



The assumptions of the one-way analysis of variance (randomness and independence, normality, and homogeneity of variance) also apply to the randomized block design. If the normality assumption is violated, you can use the Friedman rank test (see Online Section 12.9). In addition, you need to assume that there is no *interacting effect* between the groups and the blocks. In other words, you need to assume that any differences between the groups (the restaurants) are consistent across the entire set of blocks (the evaluators). The concept of *interaction* is discussed further in Section 11.3.

Did the blocking result in an increase in precision in comparing the different groups? To answer this question, you use Equation (11.14) to calculate the **estimated relative efficiency (RE)** of the randomized block design as compared with the completely randomized design.

ESTIMATED RELATIVE EFFICIENCY

$$RE = \frac{(r - 1)MSBL + r(c - 1)MSE}{(rc - 1)MSE} \quad (11.14)$$

Using Figure 11.10,

$$RE = \frac{(5)(56.675) + (6)(3)(14.986)}{(23)(14.986)} = 1.60$$

This value for relative efficiency means that it would take 1.6 times as many observations in a one-way ANOVA design as compared to the randomized block design in order to have the same precision in comparing the restaurants.

Multiple Comparisons: The Tukey Procedure

As in the case of the completely randomized design, once you reject the null hypothesis of no differences between the groups, you need to determine which groups are significantly different from the others. For the randomized block design, you can use a procedure developed by Tukey (see reference 10). Equation (11.15) gives the critical range for the **Tukey multiple comparisons procedure for randomized block designs**.

THE CRITICAL RANGE FOR THE RANDOMIZED BLOCK DESIGN

$$\text{Critical range} = Q_\alpha \sqrt{\frac{MSE}{r}} \quad (11.15)$$

where Q_α is the upper-tail critical value from a Studentized range distribution having c degrees of freedom in the numerator and $(r - 1)(c - 1)$ degrees of freedom in the denominator. Values for the Studentized range distribution are found in Table E.7.

To perform the multiple comparisons, you do the following:

1. Compute the absolute mean differences, $|\bar{X}_j - \bar{X}_{j'}|$, (where $j \neq j'$), among all $c(c - 1)/2$ pairs of sample means.
2. Compute the critical range for the Tukey procedure using Equation (11.15).
3. Compare each of the $c(c - 1)/2$ pairs against the critical range. If the absolute difference in a specific pair of sample means, say $|\bar{X}_j - \bar{X}_{j'}|$, is greater than the critical range, then group j and group j' is significantly different.
4. Interpret the results.

To apply the Tukey procedure, return to the quick-service chain study. Because there are four restaurants, there are $4(4 - 1)/2 = 6$ possible pairwise comparisons. From Figure 11.10, the absolute mean differences are

1. $|\bar{X}_{.1} - \bar{X}_{.2}| = |77.50 - 66.67| = 10.83$
2. $|\bar{X}_{.1} - \bar{X}_{.3}| = |77.50 - 91.00| = 13.50$
3. $|\bar{X}_{.1} - \bar{X}_{.4}| = |77.50 - 79.33| = 1.83$
4. $|\bar{X}_{.2} - \bar{X}_{.3}| = |66.67 - 91.00| = 24.33$
5. $|\bar{X}_{.2} - \bar{X}_{.4}| = |66.67 - 79.33| = 12.66$
6. $|\bar{X}_{.3} - \bar{X}_{.4}| = |91.00 - 79.33| = 11.67$

Locate $MSE = 14.986$ and $r = 6$ in Figure 11.10 to determine the critical range. From Table E.7 [for $\alpha = .05$, $c = 4$, and $(r - 1)(c - 1) = 15$], Q_α , the upper-tail critical value of the test statistic with 4 and 15 degrees of freedom, is 4.08. Using Equation (11.15),

$$\text{Critical range} = 4.08 \sqrt{\frac{14.986}{6}} = 6.448$$

All pairwise comparisons except $|\bar{X}_{.1} - \bar{X}_{.4}|$ are greater than the critical range. Therefore, you conclude with 95% confidence that the population mean rating is different between all pairs of restaurant branches except for Henry St. and Blvd. N. In addition, Granby has the highest mean ratings (i.e., most preferred) and Surf Ave has the lowest (i.e., least preferred).

Problems for Section 11.2

LEARNING THE BASICS

11.15 Given a randomized block experiment with five groups and seven blocks, answer the following:

- a. How many degrees of freedom are there in determining the among-group variation?
- b. How many degrees of freedom are there in determining the among-block variation?
- c. How many degrees of freedom are there in determining the random variation?
- d. How many degrees of freedom are there in determining the total variation?

11.16 From Problem 11.15,

- a. if $SSA = 60$, $SSBL = 75$, and $SST = 210$, what is SSE ?
- b. what are MSA , $MSBL$, and MSE ?
- c. what is the value of the F_{STAT} test statistic for the factor effect?
- d. what is the value of the F_{STAT} test statistic for the block effect?

11.17 From Problems 11.15 and 11.16,

- a. construct the ANOVA summary table and fill in all values in the body of the table.
- b. at the 0.05 level of significance, is there evidence of a difference in the group means?
- c. at the 0.05 level of significance, is there evidence of a difference due to blocks?

11.18 From Problems 11.15, 11.16, and 11.17,

- to perform the Tukey procedure, how many degrees of freedom are there in the numerator, and how many degrees of freedom are there in the denominator of the Studentized range distribution?
- at the 0.05 level of significance, what is the upper-tail critical value from the Studentized range distribution?
- to perform the Tukey procedure, what is the critical range?

11.19 Given a randomized block experiment with three groups and seven blocks,

- how many degrees of freedom are there in determining the among-group variation?
- how many degrees of freedom are there in determining the among-block variation?
- how many degrees of freedom are there in determining the random variation?
- how many degrees of freedom are there in determining the total variation?

11.20 From Problem 11.19, if $SSA = 36$ and the randomized block F_{STAT} statistic is 6.0,

- what are MSE and SSE ?
- what is $SSBL$ if the F_{STAT} test statistic for block effect is 4.0?
- what is SST ?
- at the 0.01 level of significance, is there evidence of an effect due to groups, and is there evidence of an effect due to blocks?

11.21 Given a randomized block experiment with four groups and eight blocks, in the following ANOVA summary table, fill in all the missing results.

Source	Degrees of Freedom	Sum of Squares	Mean Square (Variance)	F
Among groups	$c - 1 = ?$	$SSA = ?$	$MSA = 80$	$F_{STAT} = ?$
Among blocks	$r - 1 = ?$	$SSBL = 540$	$MSBL = ?$	$F_{STAT} = 5.0$
	$(r - 1)$			
Error	$(c - 1) = ?$	$SSE = ?$	$MSE = ?$	
Total	$rc - 1 = ?$	$SST = ?$		

11.22 From Problem 11.21,

- at the 0.05 level of significance, is there evidence of a difference among the four group means?
- at the 0.05 level of significance, is there evidence of an effect due to blocks?

APPLYING THE CONCEPTS

11.23 Nine experts rated four brands of Colombian coffee in a taste-testing experiment. A rating on a 7-point scale (1 = extremely unpleasing, 7 = extremely pleasing) is given for each of four characteristics: taste, aroma, richness, and acidity. The following data (stored in **Coffee**) display the summated ratings, accumulated over all four characteristics.

EXPERT	BRAND			
	A	B	C	D
C.C.	24	26	25	22
S.E.	27	27	26	24
E.G.	19	22	20	16
B.L.	24	27	25	23
C.M.	22	25	22	21
C.N.	26	27	24	24
G.N.	27	26	22	23
R.M.	25	27	24	21
P.V.	22	23	20	19

At the 0.05 level of significance, completely analyze the data to determine whether there is evidence of a difference in the summated ratings of the four brands of Colombian coffee and, if so, which of the brands are rated highest (i.e., best). What can you conclude?

11.24 How do the ratings for TV, phone, and Internet services compare? The data in **Telecom2** represent the mean ratings in 13 different cities.

Source: Data extracted from “Ratings: TV, Phone, and Internet Services,” *Consumer Reports*, May 2013, pp. 24–25.

- At the 0.05 level of significance, determine whether there is evidence of a difference in the mean rating between TV, phone, and Internet services.
- If appropriate, use the Tukey procedure to determine which services’ mean ratings differ. Again, use a 0.05 level of significance.

11.25 An article discussed the price of a market basket of items at Publix and Winn-Dixie supermarkets, and at Walmart and Super Target stores, located in South Florida. The file **Supermarket-Prices** contains the prices listed in that article (data extracted from “S. Salisbury, Supermarket Showdown,” *The Palm Beach Post*, February 11, 2011, pp. 1F, 7F)

- At the 0.05 level of significance, determine whether there is evidence of a difference in the mean prices for these items at the four supermarkets.
- What assumptions are necessary to perform this test?
- If appropriate, use the Tukey procedure to determine which supermarket mean prices differ. (Use $\alpha = 0.05$.)
- Was there a significant block effect in this experiment? Explain.

11.26 How different are the rates of return of money market accounts and certificates of deposit that vary in their term length? The data in **MMCDRate** contain the money market, 1-year CD, two-year CD, and five-year CD rates for banks in a suburban area (data extracted from “Consumer Money Rates,” *Newsday*, June 13, 2013, p. A47).

- At the 0.05 level of significance, determine whether there is evidence of a difference in the mean rates for these investments.
- What assumptions are necessary to perform this test?
- If appropriate, use the Tukey procedure to determine which investments differ. (Use $\alpha = 0.05$.)
- Was there a significant block effect in their mean rates in this experiment? Explain.

11.27 Philips Semiconductors is a leading European manufacturer of integrated circuits. Integrated circuits are produced on silicon wafers, which are ground to target thickness early in the production process. The wafers are positioned in different locations on a grinder and kept in place using vacuum decompression. One of the goals of process improvement is to reduce the variability in the thickness of the wafers in different positions and in different batches. Data were collected from a sample of 30 batches. In each batch, the thickness of the wafers on positions 1 and 2 (outer circle), 18 and 19 (middle circle), and 28 (inner circle) was measured and stored in **Circuits**. At the 0.01 level of significance, completely analyze the data to determine whether there is evidence of a difference in the mean thickness of the wafers for the five positions and, if so, which of the positions are different. What can you conclude?

Source: Data extracted from K. C. B. Roes and R. J. M. M. Does, "Shewhart-Type Charts in Nonstandard Situations," *Technometrics*, 37 (1995), pp. 15–24.

11.28 The data in **Concrete2** represent the compressive strength in thousands of pounds per square inch of 40 samples of concrete taken 2, 7, and 28 days after pouring.

Source: Data extracted from O. Carrillo-Gamboa and R. F. Gunst, "Measurement-Error-Model Collinearities," *Technometrics*, 34 (1992), pp. 454–464.

- At the 0.05 level of significance, is there evidence of a difference in the mean compressive strength after 2, 7, and 28 days?
- If appropriate, use the Tukey procedure to determine the days that differ in mean compressive strength. (Use $\alpha = 0.05$.)
- Determine the relative efficiency of the randomized block design as compared with the completely randomized (one-way ANOVA) design.
- Construct boxplots of the compressive strength for the different time periods.
- Based on the results of (a), (b), and (d), is there a pattern in the compressive strength over the three time periods?

11.3 The Factorial Design: Two-Way ANOVA

In Section 11.1, you learned about the completely randomized design and in Section 11.2, you learned about the randomized block design. In this section, the single-factor completely randomized design is extended to the **two-factor factorial design**, in which two factors are simultaneously evaluated. Each factor is evaluated at two or more levels. For example, in the Arlington's scenario on page 394, the company faces the business problem of simultaneously evaluating four locations and the effectiveness of providing mobile payment to determine which location should be used and whether mobile payment should be made available. Although this section uses only two factors, you can extend factorial designs to three or more factors (see references 4, 5, 6, 7, and 10).

To analyze data from a two-factor factorial design, you use **two-way ANOVA**. The following definitions are needed to develop the two-way ANOVA procedure:

r = number of levels of factor A

c = number of levels of factor B

n' = number of values (replicates) for each cell (combination of a particular level of factor A and a particular level of factor B)

n = number of values in the entire experiment (where $n = rcn'$)

X_{ijk} = value of the k th observation for level i of factor A and level j of factor B

$$\bar{X} = \frac{\sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n'} X_{ijk}}{rcn'} = \text{grand mean}$$

$$\bar{X}_{i..} = \frac{\sum_{j=1}^c \sum_{k=1}^{n'} X_{ijk}}{cn'} = \text{mean of the } i\text{th level of factor A (where } i = 1, 2, \dots, r\text{)}$$

$$\bar{X}_{.j} = \frac{\sum_{i=1}^r \sum_{k=1}^{n'} X_{ijk}}{rn'} = \text{mean of the } j\text{th level of factor B (where } i = 1, 2, \dots, c\text{)}$$

$$\bar{X}_{ij.} = \frac{\sum_{k=1}^{n'} X_{ijk}}{n'} = \text{mean of the cell } ij, \text{ the combination of the } i\text{th level of factor A and the } j\text{th level of factor B}$$

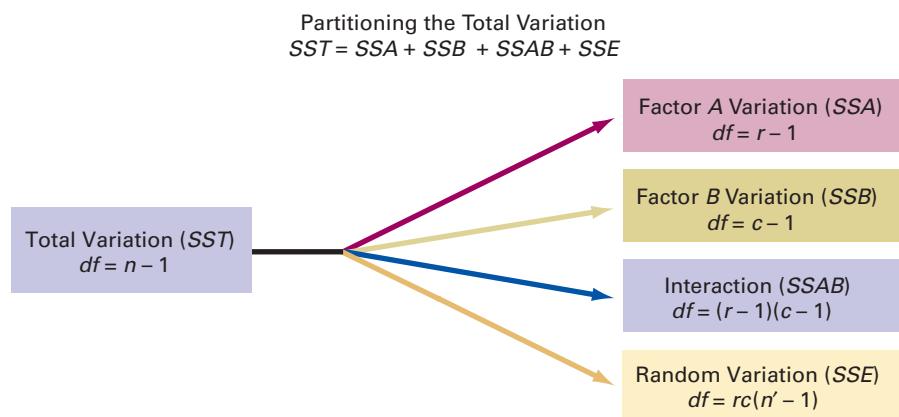
Because of the complexity of these computations, you should only use computerized methods when performing this analysis. However, to help explain the two-way ANOVA, the decomposition of the total variation is illustrated. In this discussion, only cases in which there are an equal number of values (also called **replicates**) (sample sizes n') for each combination of the levels of factor A with those of factor B are considered. (See references 1 and 6 for a discussion of two-factor factorial designs with unequal sample sizes.)

Factor and Interaction Effects

There is an **interaction** between factors A and B if the effect of factor A is different for various levels of factor B. Thus, when dividing the total variation into different sources of variation, you need to account for a possible interaction effect, as well as for factor A, factor B, and random error. To accomplish this, the total variation (SST) is subdivided into sum of squares due to factor A (or SSA), sum of squares due to factor B (or SSB), sum of squares due to the interaction effect of A and B (or $SSAB$), and sum of squares due to random variation (or SSE). This decomposition of the total variation (SST) is displayed in Figure 11.13.

FIGURE 11.13

Partitioning the total variation in a two-factor factorial design



The sum of squares total (SST) represents the total variation among all the values around the grand mean. Equation (11.16) shows the computation for total variation.

TOTAL VARIATION IN TWO-WAY ANOVA

$$SST = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n'} (X_{ijk} - \bar{X})^2 \quad (11.16)$$

The **sum of squares due to factor A (SSA)** represents the differences among the various levels of factor A and the grand mean. Equation (11.17) shows the computation for factor A variation.

FACTOR A VARIATION

$$SSA = cn' \sum_{i=1}^r (\bar{X}_{i..} - \bar{\bar{X}})^2 \quad (11.17)$$

The **sum of squares due to factor B (SSB)** represents the differences among the various levels of factor B and the grand mean. Equation (11.18) shows the computation for factor B variation.

FACTOR B VARIATION

$$SSB = rn' \sum_{j=1}^c (\bar{X}_{j\cdot} - \bar{\bar{X}})^2 \quad (11.18)$$

The **sum of squares due to interaction (SSAB)** represents the interacting effect of specific combinations of factor A and factor B . Equation (11.19) shows the computation for interaction variation.

INTERACTION VARIATION

$$SSAB = n' \sum_{i=1}^r \sum_{j=1}^c (\bar{X}_{ij\cdot} - \bar{X}_{i\cdot} - \bar{X}_{j\cdot} + \bar{\bar{X}})^2 \quad (11.19)$$

The **sum of squares error (SSE)** represents random variation—that is, the differences among the values within each cell and the corresponding cell mean. Equation (11.20) shows the computation for random variation.

RANDOM VARIATION IN TWO-WAY ANOVA

$$SSE = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n'} (X_{ijk} - \bar{X}_{ij\cdot})^2 \quad (11.20)$$

Because there are r levels of factor A , there are $r - 1$ degrees of freedom associated with SSA . Similarly, because there are c levels of factor B , there are $c - 1$ degrees of freedom associated with SSB . Because there are n' replicates in each of the rc cells, there are $rc(n' - 1)$ degrees of freedom associated with the SSE term. Carrying this further, there are $n - 1$ degrees of freedom associated with the sum of squares total (SST) because you are comparing each value, X_{ijk} , to the grand mean, $\bar{\bar{X}}$, based on all n values. Therefore, because the degrees of freedom for each of the sources of variation must add to the degrees of freedom for the total variation (SST), you can calculate the degrees of freedom for the interaction component ($SSAB$) by subtraction. The degrees of freedom for interaction are $(r - 1)(c - 1)$.

If you divide each sum of squares by its associated degrees of freedom, you have the four variances or mean square terms (MSA , MSB , $MSAB$, and MSE). Equations (11.21a-d) give the mean square terms needed for the two-way ANOVA table.

MEAN SQUARES IN TWO-WAY ANOVA

$$MSA = \frac{SSA}{r - 1} \quad (11.21a)$$

$$MSB = \frac{SSB}{c - 1} \quad (11.21b)$$

Student Tip

Remember, *mean square* is another term for *variance*.

$$MSAB = \frac{SSAB}{(r-1)(c-1)} \quad (11.21c)$$

$$MSE = \frac{SSE}{rc(n'-1)} \quad (11.21d)$$

Testing for Factor and Interaction Effects

There are three different tests to perform in a two-way ANOVA:

- A test of the hypothesis of no difference due to factor A
- A test of the hypothesis of no difference due to factor B
- A test of the hypothesis of no interaction of factors A and B

To test the hypothesis of no difference due to factor A :

$$H_0: \mu_{1..} = \mu_{2..} = \dots = \mu_{r..}$$

against the alternative:

$$H_1: \text{Not all } \mu_{i..} \text{ are equal}$$

you use the F_{STAT} test statistic in Equation (11.22).

F TEST FOR FACTOR A EFFECT

$$F_{STAT} = \frac{MSA}{MSE} \quad (11.22)$$

You reject the null hypothesis at the α level of significance if

$$F_{STAT} = \frac{MSA}{MSE} > F_\alpha$$

where F_α is the upper-tail critical value from an F distribution with $r-1$ and $rc(n'-1)$ degrees of freedom.

To test the hypothesis of no difference due to factor B :

$$H_0: \mu_{.1} = \mu_{.2} = \dots = \mu_{.c}$$

against the alternative:

$$H_1: \text{Not all } \mu_{.j} \text{ are equal}$$

you use the F_{STAT} test statistic in Equation (11.23).

F TEST FOR FACTOR B EFFECT

$$F_{STAT} = \frac{MSB}{MSE} \quad (11.23)$$

You reject the null hypothesis at the α level of significance if

$$F_{STAT} = \frac{MSB}{MSE} > F_\alpha$$

where F_α is the upper-tail critical value from an F distribution with $c - 1$ and $rc(n' - 1)$ degrees of freedom.

To test the hypothesis of no interaction of factors A and B :

H_0 : The interaction of A and B is equal to zero

against the alternative:

H_1 : The interaction of A and B is not equal to zero

you use the F_{STAT} test statistic in Equation (11.24).

F TEST FOR INTERACTION EFFECT

$$F_{STAT} = \frac{MSAB}{MSE} \quad (11.24)$$

You reject the null hypothesis at the α level of significance if

Student Tip

In each of these F tests, the denominator of the F_{STAT} statistic is MSE .

where F_α is the upper-tail critical value from an F distribution with $(r - 1)(c - 1)$ and $rc(n' - 1)$ degrees of freedom.

Table 11.8 presents the entire two-way ANOVA table.

TABLE 11.8

Analysis of Variance Table for the Two-Factor Factorial Design

Source	Degrees of Freedom	Sum of Squares	Mean Square (Variance)	F
A	$r - 1$	SSA	$MSA = \frac{SSA}{r - 1}$	$F_{STAT} = \frac{MSA}{MSE}$
B	$c - 1$	SSB	$MSB = \frac{SSB}{c - 1}$	$F_{STAT} = \frac{MSB}{MSE}$
AB	$(r - 1)(c - 1)$	$SSAB$	$MSAB = \frac{SSAB}{(r - 1)(c - 1)}$	$F_{STAT} = \frac{MSAB}{MSE}$
Error	$rc(n' - 1)$	SSE	$MSE = \frac{SSE}{rc(n' - 1)}$	
Total	$n - 1$	SST		

To illustrate two-way ANOVA, return to the Arlington's scenario on page 394. As a member of the sales team, you first explored how different in-store locations might affect the sales of mobile electronics merchandise using one-way ANOVA. Now, to explore the effects of permitting mobile payment methods to buy mobile electronics merchandise, you can design an experiment that examines this second (B) factor as it studies the effects of in-store location (factor A) using two-way ANOVA. Two-way ANOVA will allow you to determine if there is a significant difference in mobile electronics sales among the four in-store locations *and* whether permitting mobile payment methods makes a difference.

To test the effects of the two factors, you conduct a 60-day experiment at 40 same-sized stores that have similar storewide net sales. You randomly assign ten stores to use the current in-aisle location, ten stores to use the front of the store near weekly specials, ten stores to use the end-cap special kiosk display, and ten stores to use the location adjacent to the Expert Counter. In five stores in each of the four groups, you permit mobile payment methods

(for the other five in each group, mobile payment methods are not permitted). At the end of the experiment, you organize the mobile electronics sales data by group and store the data in **Mobile Electronics2**. Table 11.9 presents the data of the experiment.

TABLE 11.9

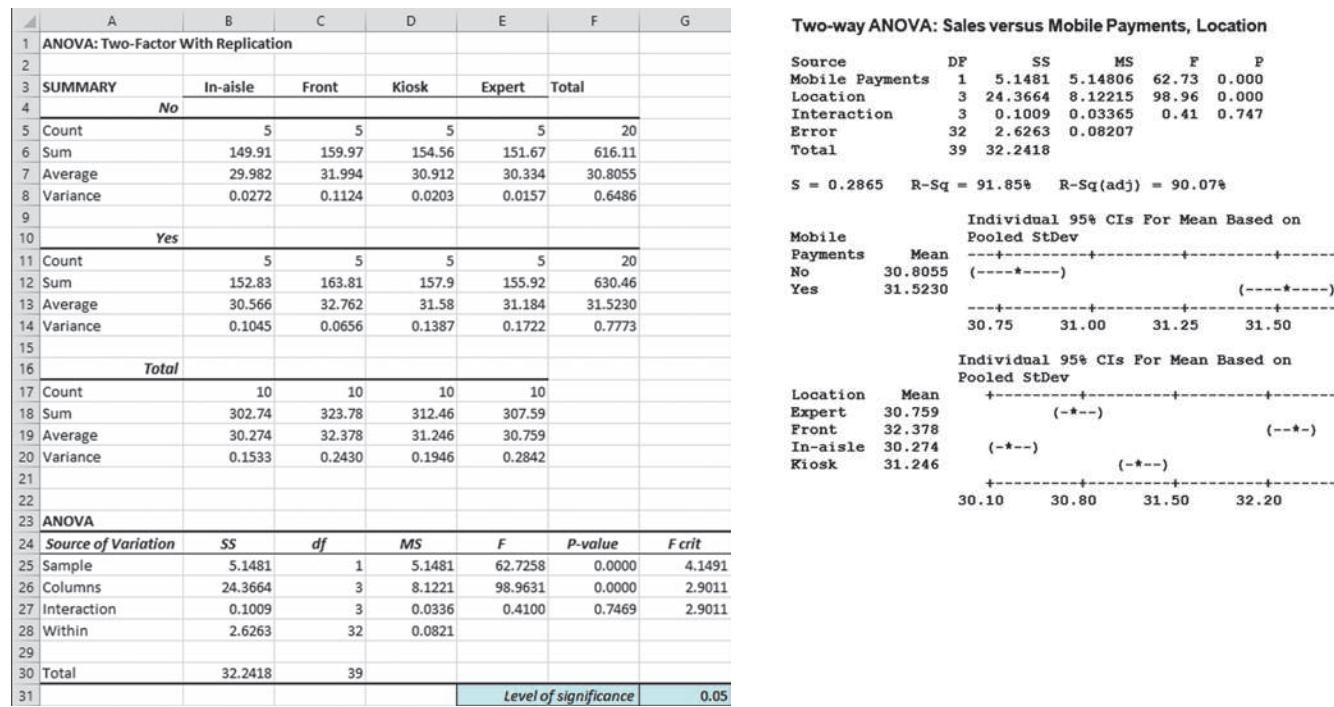
Mobile Electronics Sales at Four In-Store Locations with Mobile Payments Permitted and Not Permitted

MOBILE PAYMENTS	IN-STORE LOCATION			
	In-Aisle	Front	Kiosk	Expert
No	30.06	32.22	30.78	30.33
No	29.96	31.47	30.91	30.29
No	30.19	32.13	30.79	30.25
No	29.96	31.86	30.95	30.25
No	29.74	32.29	31.13	30.55
Yes	30.66	32.81	31.34	31.03
Yes	29.99	32.65	31.80	31.77
Yes	30.73	32.81	32.00	30.97
Yes	30.72	32.42	31.07	31.43
Yes	30.73	33.12	31.69	30.72

Figure 11.14 presents the results for this example. In the Excel results, the *A*, *B*, and Error sources of variation in Table 11.8 above are labeled Sample, Columns, and Within, respectively. In the Minitab results, the names of the factors (Mobile Payments and In-Store Location) are used to label the *A* and *B* sources of variation.

FIGURE 11.14

Excel and Minitab two-way ANOVA results for the in-store location sales and mobile payment experiment



To interpret the results, you start by testing whether there is an interaction effect between factor *A* (mobile payments) and factor *B* (in-store locations). If the interaction effect is significant, further analysis will focus on this interaction. If the interaction effect is not significant, you can focus on the **main effects**—the potential effect of permitting mobile payment (factor *A*) and the potential differences in in-store locations (factor *B*).

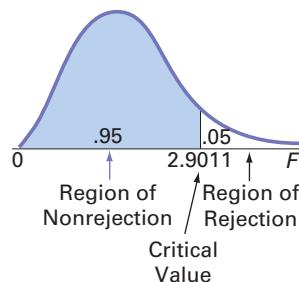
¹Table E.5 does not provide the upper-tail critical values from the F distribution with 32 degrees of freedom in the denominator. When the desired degrees of freedom are not provided in the table, use the p -value computed by Excel or Minitab.

Using the 0.05 level of significance, to determine whether there is evidence of an interaction effect, you reject the null hypothesis of no interaction between mobile payments and in-store locations if the computed F_{STAT} statistic is greater than 2.9011, the upper-tail critical value from the F distribution, with 3 and 32 degrees of freedom (see Figures 11.14 and 11.15).¹

Because $F_{STAT} = 0.4100 < 2.9011$ or the p -value = 0.7469 > 0.05, you do not reject H_0 . You conclude that there is insufficient evidence of an interaction effect between mobile payment and in-store location. You can now focus on the main effects.

FIGURE 11.15

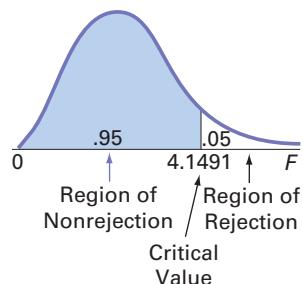
Regions of rejection and nonrejection at the 0.05 level of significance, with 3 and 32 degrees of freedom



Using the 0.05 level of significance and testing whether there is an effect due to mobile payment options (yes or no) (factor A), you reject the null hypothesis if the computed F_{STAT} test statistic is greater than 4.1491, the upper-tail critical value from the F distribution with 1 and 32 degrees of freedom (see Figures 11.14 and 11.16). Because $F_{STAT} = 62.7258 > 4.1491$ or the p -value = 0.0000 < 0.05, you reject H_0 . You conclude that there is evidence of a difference in the mean sales when mobile payment methods are permitted as compared to when they are not. Because the mean sales when mobile payment methods are permitted is 31.523 and is 30.806 when they are not, you can conclude that permitting mobile payment methods has led to an increase in mean sales.

FIGURE 11.16

Regions of rejection and nonrejection at the 0.05 level of significance, with 1 and 32 degrees of freedom



Using the 0.05 level of significance and testing for a difference among the in-store locations (factor B), you reject the null hypothesis of no difference if the computed F_{STAT} test statistic is greater than 2.9011, the upper-tail critical value from the F distribution with 3 degrees of freedom in the numerator and 32 degrees of freedom in the denominator (see Figures 11.14 and 11.15). Because $F_{STAT} = 98.9631 > 2.9011$ or the p -value = 0.0000 < 0.05, you reject H_0 . You conclude that there is evidence of a difference in the mean sales among the four in-store locations.

Multiple Comparisons: The Tukey Procedure

If one or both of the factor effects are significant and there is no significant interaction effect, when there are more than two levels of a factor, you can determine the particular levels that are significantly different by using the **Tukey multiple comparisons procedure for two-way ANOVA** (see references 6 and 10). Equation (11.25) gives the critical range for factor A .

CRITICAL RANGE FOR FACTOR A

$$\text{Critical range} = Q_\alpha \sqrt{\frac{MSE}{cn'}} \quad (11.25)$$

where Q_α is the upper-tail critical value from a Studentized range distribution having r and $rc(n' - 1)$ degrees of freedom. (Values for the Studentized range distribution are found in Table E.7.)

Equation (11.26) gives the critical range for factor B .

CRITICAL RANGE FOR FACTOR B

$$\text{Critical range} = Q_\alpha \sqrt{\frac{MSE}{rn'}} \quad (11.26)$$

where Q_α is the upper-tail critical value from a Studentized range distribution having c and $rc(n' - 1)$ degrees of freedom. (Values for the Studentized range distribution are found in Table E.7.)

To use the Tukey procedure, return to the mobile electronics sales data of Table 11.7 on page 423. In the ANOVA summary table in Figure 11.14 on page 423, the interaction effect is not significant. Because there are only two categories for mobile payment (yes and no), there are no multiple comparisons to be constructed. Using $\alpha = 0.05$, there is evidence of a significant difference among the four in-store locations that comprise factor B . Thus, you can use the Tukey multiple comparisons procedure to determine which of the four in-store locations differ.

Because there are four in-store locations, there are $4(4 - 1)/2 = 6$ pairwise comparisons. Using the calculations presented in Figure 11.14, the absolute mean differences are as follows:

1. $|\bar{X}_{.1.} - \bar{X}_{.2.}| = |30.274 - 32.378| = 2.104$
2. $|\bar{X}_{.1.} - \bar{X}_{.3.}| = |30.274 - 31.246| = 0.972$
3. $|\bar{X}_{.1.} - \bar{X}_{.4.}| = |30.274 - 30.759| = 0.485$
4. $|\bar{X}_{.2.} - \bar{X}_{.3.}| = |32.378 - 31.246| = 1.132$
5. $|\bar{X}_{.2.} - \bar{X}_{.4.}| = |32.378 - 30.759| = 1.619$
6. $|\bar{X}_{.3.} - \bar{X}_{.4.}| = |31.246 - 30.759| = 0.487$

To determine the critical range, refer to Figure 11.14 to find $MSE = 0.0821$, $r = 2$, $c = 4$, and $n' = 5$. From Table E.7 [for $\alpha = 0.05$, $c = 4$, and $rc(n' - 1) = 32$], Q_α , the upper-tail critical value of the Studentized range distribution with 4 and 32 degrees of freedom is approximately 3.84. Using Equation (11.17),

$$\text{Critical range} = 3.84 \sqrt{\frac{0.0821}{10}} = 0.3482$$

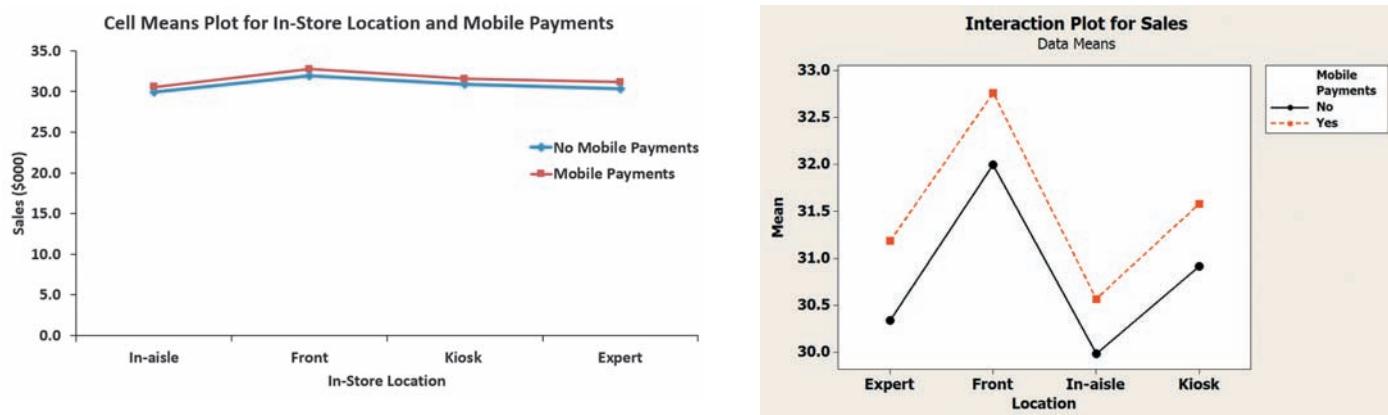
Because each of the six comparisons is greater than the critical range of 0.3482, you can conclude that the population mean sales is different for the four in-store locations. The front location is estimated to have higher mean sales than the other three in-store locations. The kiosk location is estimated to have higher mean sales than the in-aisle and expert locations. The expert location is estimated to have higher mean sales than the in-aisle location. Note that by using $\alpha = 0.05$, you are able to make all six comparisons with an overall error rate of only 5%.

Visualizing Interaction Effects: The Cell Means Plot

You can get a better understanding of the interaction effect by plotting the **cell means**, the means of all possible factor-level combinations. Figure 11.17 presents a cell means plot that uses the cell means for the mobile payments permitted/in-store location combinations shown in Figure 11.14 on page 423. From the plot of the mean sales for each combination of mobile payments permitted and in-store location, observe that the two lines (representing the two levels of mobile payments, yes and no,) are roughly parallel. This indicates that the *difference* between the mean sales for stores that permit mobile payment methods and those that do not is virtually the same for the four in-store locations. In other words, there is no *interaction* between these two factors, as was indicated by the *F* test.

FIGURE 11.17

Excel and Minitab cell means plots for mobile electronic sales based on mobile payments permitted and in-store location



Interpreting Interaction Effects

How do you interpret an interaction? When there is an interaction, some levels of factor *A* respond better with certain levels of factor *B*. For example, with respect to mobile electronics sales, suppose that some in-store locations were better when mobile payment methods were permitted and other in-store locations were better when mobile payment methods were not permitted. If this were true, the lines of Figure 11.17 would not be nearly as parallel, and the interaction effect might be statistically significant. In such a situation, the difference between whether mobile payment methods were permitted is no longer the same for all in-store locations. Such an outcome would also complicate the interpretation of the *main effects* because differences in one factor (whether mobile payment methods were permitted) would not be consistent across the other factor (the in-store locations).

Example 11.2 illustrates a situation with a significant interaction effect.

EXAMPLE 11.2

Interpreting Significant Interaction Effects

A nationwide company specializing in preparing students for college and graduate school entrance exams, such as the SAT, ACT, GRE, and LSAT, had the business objective of improving its ACT preparatory course. Two factors of interest to the company are the length of the course (a condensed 10-day period or a regular 30-day period) and the type of course (traditional classroom or online distance learning). The company collected data by randomly assigning 10 clients to each of the four cells that represent a combination of length of the course and type of course. The results are organized in the file **ACT** and presented in Table 11.10.

What are the effects of the type of course and the length of the course on ACT scores?

TABLE 11.10

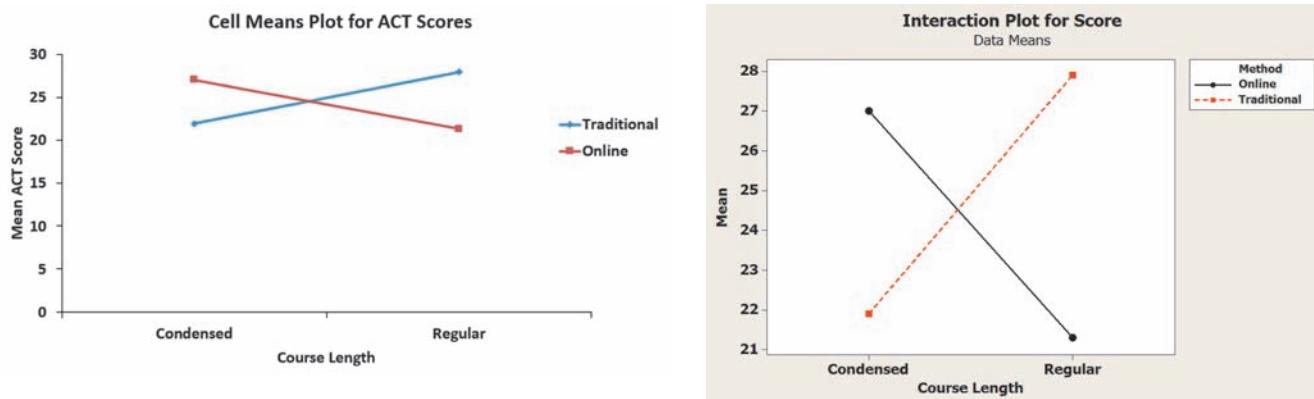
ACT Scores for Different Types and Lengths of Courses

TYPE OF COURSE	LENGTH OF COURSE			
	Condensed		Regular	
Traditional	26	18	34	28
Traditional	27	24	24	21
Traditional	25	19	35	23
Traditional	21	20	31	29
Traditional	21	18	28	26
Online	27	21	24	21
Online	29	32	16	19
Online	30	20	22	19
Online	24	28	20	24
Online	30	29	23	25

SOLUTION The cell means plot presented in Figure 11.18 shows a strong interaction between the type of course and the length of the course. The nonparallel lines indicate that the effect of condensing the course depends on whether the course is taught in the traditional classroom or by online distance learning. The online mean score is higher when the course is condensed to a 10-day period, whereas the traditional mean score is higher when the course takes place over the regular 30-day period.

FIGURE 11.18

Excel and Minitab cell means plot of ACT scores



To verify the visual analysis provided by interpreting the cell means plot, you begin by testing whether there is a statistically significant interaction between factor *A* (length of course) and factor *B* (type of course). Using a 0.05 level of significance, you reject the null hypothesis because $F_{STAT} = 24.2569 > 4.1132$ or the *p*-value equals $0.0000 < 0.05$ (see Figure 11.19 on next page). Thus, the hypothesis test confirms the interaction evident in the cell means plot.

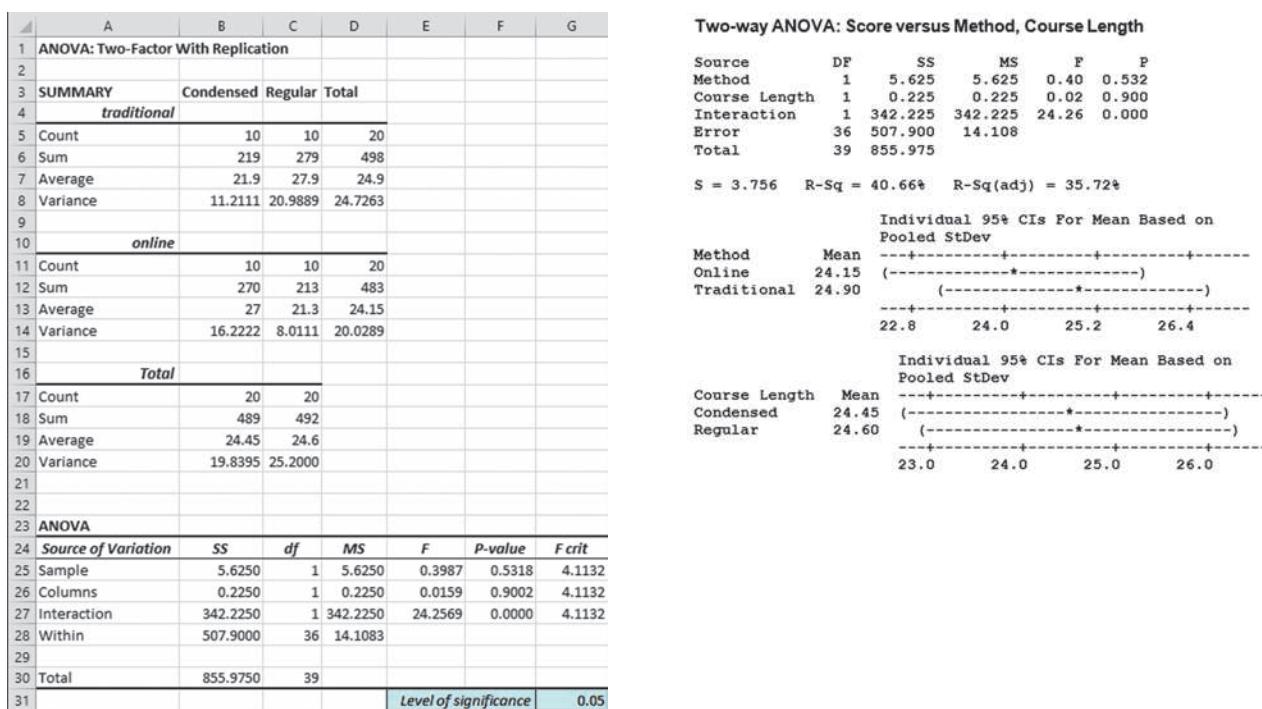
The existence of this significant interaction effect complicates the interpretation of the hypothesis tests concerning the two main effects. You cannot directly conclude that there is no effect with respect to length of course and type of course, even though both have *p*-values > 0.05 .

Given that the interaction is significant, you can reanalyze the data with the two factors collapsed into four groups of a single factor rather than a two-way ANOVA with two levels of each of the two factors. You can reorganize the data as follows: Group 1 is traditional

(continued)

FIGURE 11.19

Excel and Minitab two-way ANOVA results for the ACT scores



condensed, Group 2 is traditional regular, Group 3 is online condensed, and Group 4 is online regular. Figure 11.20 shows the results for these data, stored in **ACT-OneWay**.

From Figure 11.20, because $F_{STAT} = 8.2239 > 2.8663$ or $p\text{-value} = 0.0003 < 0.05$, there is evidence of a significant difference in the four groups (traditional condensed, traditional regular, online condensed, and online regular). Traditional condensed is different from traditional regular and from online condensed. Traditional regular is also different from online regular, and online condensed is also different from online regular. Thus, whether condensing a course is a good idea depends on whether the course is offered in a traditional classroom or as an online distance learning course. To ensure the highest mean ACT scores, the company should use the traditional approach for courses that are given over a 30-day period but use the online approach for courses that are condensed into a 10-day period.

FIGURE 11.20

Excel (below) and Minitab (page 429) one-way ANOVA and Tukey-Kramer results for the ACT scores

A	B	C	D	E	F	G
1 One-Way ANOVA (ANOVA: Single Factor)						
2						
3 SUMMARY						
4 Groups						
5 Group 1	10	219	21.9	11.2111		
6 Group 2	10	279	27.9	20.9889		
7 Group 3	10	270	27	16.2222		
8 Group 4	10	213	21.3	8.0111		
9						
10						
11 ANOVA						
12 Source of Variation	SS	df	MS	F	P-value	F crit
13 Between Groups	348.0750	3	116.0250	8.2239	0.0003	2.8663
14 Within Groups	507.9000	36	14.1083			
15						
16 Total	855.9750	39				
17					Level of significance	0.05

A	B	C	D	E	F	G	H	I
1 Tukey Kramer Multiple Comparisons								
2								
3								
4 Group	Sample Mean	Sample Size		Absolute Difference	Std. Error of Difference	Critical Range	Results	
5 1: Group 1	21.9	10	Group 1 to Group 2	6	1.1878	4.8105	Means are different	
6 2: Group 2	27.9	10	Group 1 to Group 3	5.1	1.1878	4.8105	Means are different	
7 3: Group 3	27	10	Group 1 to Group 4	0.6	1.1878	4.8105	Means are not different	
8 4: Group 4	21.3	10	Group 2 to Group 3	0.9	1.1878	4.8105	Means are not different	
9			Group 2 to Group 4	6.6	1.1878	4.8105	Means are different	
10 Other Data			Group 3 to Group 4	5.7	1.1878	4.8105	Means are different	
11 Level of significance	0.05							
12 Numerator d.f.	4							
13 Denominator d.f.	36							
14 MSW	14.1083							
15 Q Statistic	4.05							

One-way ANOVA: Group 1, Group 2, Group 3, Group 4

Source	DF	SS	MS	F	P
Factor	3	348.1	116.0	8.22	0.000
Error	36	507.9	14.1		
Total	39	856.0			

S = 3.756 R-Sq = 40.66% R-Sq(adj) = 35.72%

Individual 95% CIs For Mean Based on Pooled StDev

Level	N	Mean	StDev	-----+-----+-----+-----+
Group 1	10	21.900	3.348	(-----*-----)
Group 2	10	27.900	4.581	(-----*-----)
Group 3	10	27.000	4.028	(-----*-----)
Group 4	10	21.300	2.830	(-----*-----)

21.0 24.0 27.0 30.0

Pooled StDev = 3.756

Grouping Information Using Tukey Method

	N	Mean	Grouping
Group 2	10	27.900	A
Group 3	10	27.000	A
Group 1	10	21.900	B
Group 4	10	21.300	B

Means that do not share a letter are significantly different.

**Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons**

Individual confidence level = 98.93%

Group 1 subtracted from:

	Lower	Center	Upper	-----+-----+-----+-----+
Group 2	1.475	6.000	10.525	(-----*-----)
Group 3	0.575	5.100	9.625	(-----*-----)
Group 4	-5.125	-0.600	3.925	(-----*-----)

-6.0 0.0 6.0 12.0

Group 2 subtracted from:

	Lower	Center	Upper	-----+-----+-----+-----+
Group 3	-5.425	-0.900	3.625	(-----*-----)
Group 4	-11.125	-6.600	-2.075	(-----*-----)

-6.0 0.0 6.0 12.0

Group 3 subtracted from:

	Lower	Center	Upper	-----+-----+-----+-----+
Group 4	-10.225	-5.700	-1.175	(-----*-----)

-6.0 0.0 6.0 12.0

Problems for Section 11.3

LEARNING THE BASICS

11.29 Consider a two-factor factorial design with three levels for factor A, three levels for factor B, and four replicates in each of the nine cells.

- How many degrees of freedom are there in determining the factor A variation and the factor B variation?
- How many degrees of freedom are there in determining the interaction variation?
- How many degrees of freedom are there in determining the random variation?
- How many degrees of freedom are there in determining the total variation?

11.30 Assume that you are working with the results from Problem 11.29, and $SSA = 120$, $SSB = 110$, $SSE = 270$, and $SST = 540$.

- What is $SSAB$?
- What are MSA and MSB ?
- What is $MSAB$?
- What is MSE ?

11.31 Assume that you are working with the results from Problems 11.29 and 11.30.

- What is the value of the F_{STAT} test statistic for the interaction effect?
- What is the value of the F_{STAT} test statistic for the factor A effect?
- What is the value of the F_{STAT} test statistic for the factor B effect?
- Form the ANOVA summary table and fill in all values in the body of the table.

11.32 Given the results from Problems 11.29 through 11.31,

- at the 0.05 level of significance, is there an effect due to factor A?

- at the 0.05 level of significance, is there an effect due to factor B?

11.33 Given a two-way ANOVA with two levels for factor A, five levels for factor B, and four replicates in each of the 10 cells, with $SSA = 18$, $SSB = 64$, $SSE = 60$, and $SST = 150$,

- form the ANOVA summary table and fill in all values in the body of the table.
- at the 0.05 level of significance, is there an effect due to factor A?
- at the 0.05 level of significance, is there an effect due to factor B?
- at the 0.05 level of significance, is there an interaction effect?

11.34 Given a two-factor factorial experiment and the ANOVA summary table that follows, fill in all the missing results:

Source	Degrees of Freedom	Sum of Squares	Mean Square (Variance)		F
			MSA	MSB	
A	$r - 1 = 2$	$SSA = ?$	80	$F_{STAT} = ?$	
B	$c - 1 = ?$	$SSB = 220$	$MSB = ?$	$F_{STAT} = 11.0$	
AB	$(r - 1)(c - 1) = 8$	$SSAB = ?$	$MSAB = 10$	$F_{STAT} = ?$	
Error	$rc(n' - 1) = 30$	$SSE = ?$	$MSE = ?$		
Total	$n - 1 = ?$	$SST = ?$			

11.35 Given the results from Problem 11.34,

- at the 0.05 level of significance, is there an effect due to factor A?
- at the 0.05 level of significance, is there an effect due to factor B?
- at the 0.05 level of significance, is there an interaction effect?

APPLYING THE CONCEPTS

11.36 An experiment was conducted to study the extrusion process of biodegradable packaging foam. Two of the factors considered for their effect on the unit density (mg/ml) were the die temperature (145°C vs. 155°C) and the die diameter (3 mm vs. 4 mm). The results are stored in **PackagingFoam1**. (Data extracted from W. Y. Koh, K. M. Eskridge, and M. A. Hanna, “Supersaturated Split-Plot Designs,” *Journal of Quality Technology*, 45, January 2013, pp. 61–72.)

At the 0.05 level of significance,

- is there an interaction between die temperature and die diameter?
- is there an effect due to die temperature?
- is there an effect due to die diameter?
- Plot the mean unit density for each die temperature for each die diameter.
- What can you conclude about the effect of die temperature and die diameter on mean unit density?

11.37 Referring to Problem 11.36, the effect of die temperature and die diameter on the foam diameter was also measured and the results stored in **PackagingFoam2**.

At the 0.05 level of significance,

- is there an interaction between die temperature and die diameter?
- is there an effect due to die temperature?
- is there an effect due to die diameter?
- Plot the mean foam diameter for each die temperature and die diameter.
- What conclusions can you reach concerning the importance of each of these two factors on the foam diameter?

 **SELF TEST** **11.38** A student team in a business statistics course performed a factorial experiment to investigate the time required for pain-relief tablets to dissolve in a glass of water. The two factors of interest were brand name (Equate, Kroger, or Alka-Seltzer) and water temperature (hot or cold). The experiment consisted of four replicates for each of the six factor combinations. The following data, stored in **PainRelief**, show the time a tablet took to dissolve (in seconds) for the 24 tablets used in the experiment:

PAIN-RELIEF TABLET BRAND			
WATER	Equate	Kroger	Alka-Seltzer
Cold	85.87	75.98	100.11
Cold	78.69	87.66	99.65
Cold	76.42	85.71	100.83
Cold	74.43	86.31	94.16
Hot	21.53	24.10	23.80
Hot	26.26	25.83	21.29
Hot	24.95	26.32	20.82
Hot	21.52	22.91	23.21

At the 0.05 level of significance,

- is there an interaction between brand of pain reliever and water temperature?
- is there an effect due to brand?
- is there an effect due to water temperature?
- Plot the mean dissolving time for each brand for each water temperature.
- Discuss the results of (a) through (d).

11.39 A metallurgy company wanted to investigate the effect of the percentage of ammonium and the stir rate on the density of the powder produced. The results (stored in **Density**) are as follows:

AMMONIUM (%)	STIR RATE	
	100	150
2	10.95	7.54
2	14.68	6.66
2	17.68	8.03
2	15.18	8.84
30	12.65	12.46
30	15.12	14.96
30	17.48	14.96
30	15.96	12.62

Source: Extracted from L. Johnson and K. McNeilly, “Results May Not Vary,” *Quality Progress*, May 2011, pp. 41–48.

At the 0.05 level of significance,

- is there an interaction between the percentage of ammonium and the stir rate?
- is there an effect due to the percentage of ammonium?
- is there an effect due to the stir rate?
- Plot the mean density for each percentage of ammonium for each stir rate.
- Discuss the results of (a) through (d).

11.40 An experiment was conducted to try to resolve a problem of brake discs overheating at high speed on construction equipment. Five different brake discs were measured by two different temperature gauges. The temperature of each brake disc and gauge combination was measured at eight different times and the results stored in **Brakes**.

Source: Data extracted from M. Awad, T. P. Erdmann, V. Shansal, and B. Barth, “A Measurement System Analysis Approach for Hard-to-Repeat Events,” *Quality Engineering* 21 (2009): 300–305.

At the 0.05 level of significance,

- is there an interaction between the brake discs and the gauges?
- is there an effect due to brake discs?
- is there an effect due to the gauges?
- Plot the mean temperature for each brake disc for each gauge.
- Discuss the results of (a) through (d).

11.4 Fixed Effects, Random Effects, and Mixed Effects Models

Sections 11.1 through 11.3 do not consider the distinction between how the levels of a factor were selected. The equation for the F test depends on whether the levels of a factor were specifically selected or randomly selected from a population. The **Section 11.4 online topic** presents the appropriate F tests to use when the levels of a factor are either specifically selected or randomly selected from a population of levels.

USING STATISTICS

The Means to Find Differences at Arlington's Revisited

In the Arlington's scenario, you needed to determine whether there were differences in mobile electronics sales among four in-store locations as well as determine whether permitting mobile payments had an effect on those sales.

Using the one-way ANOVA, you determined that there was a difference in the mean sales for the four in-store locations. You then were able to conclude that the mean sales for the front location was higher than the current in-aisle or experimental kiosk or expert locations, that the kiosk location mean sales were higher than the in-aisle or expert locations, and that there was no evidence of a difference between the mean sales for the in-aisle and expert locations. Using the two-way ANOVA, you determined that there was no interaction between in-store location and permitting mobile payment methods and that mean sales were higher when mobile payment methods were permitted than when such methods

were not. In addition, you concluded that the population mean sales is different for the four in-store locations and reached these other conclusions:

- The front location is estimated to have higher mean sales than the other three locations.
- The kiosk location is estimated to have higher mean sales than the current in-aisle location or expert location.
- The expert location is estimated to have higher mean sales than the current in-aisle location.

Your next step as a member of the sales team might be to further investigate the differences among the sales locations as well as examine other factors that could influence mobile electronics sale.



Pavel L Photo and Video/Shutterstock

SUMMARY

In this chapter, various statistical procedures were used to analyze the effect of one or two factors of interest. The assumptions required for using these procedures were discussed in detail. Remember that you need to critically

investigate the validity of the assumptions underlying the hypothesis-testing procedures. Table 11.11 summarizes the topics covered in this chapter.

TABLE 11.11

Summary of Topics in Chapter 11

Type of Analysis (numerical data only)	Number of Factors
Comparing more than two groups	One-way analysis of variance (Section 11.1) Randomized block design (Section 11.2) Two-way analysis of variance (Section 11.3)

REFERENCES

1. Berenson, M. L., D. M. Levine, and M. Goldstein. *Intermediate Statistical Methods and Applications: A Computer Package Approach*. Upper Saddle River, NJ: Prentice Hall, 1983.
2. Conover, W. J. *Practical Nonparametric Statistics*, 3rd ed. New York: Wiley, 2000.
3. Daniel, W. W. *Applied Nonparametric Statistics*, 2nd ed. Boston: PWS Kent, 1990.
4. Gitlow, H. S., and D. M. Levine. *Six Sigma for Green Belts and Champions: Foundations, DMAIC, Tools, Cases, and Certification*. Upper Saddle River, NJ: Financial Times/Prentice Hall, 2005.
5. Hicks, C. R., and K. V. Turner. *Fundamental Concepts in the Design of Experiments*, 5th ed. New York: Oxford University Press, 1999.
6. Kutner, M. H., J. Neter, C. Nachtsheim, and W. Li. *Applied Linear Statistical Models*, 5th ed. New York: McGraw-Hill-Irwin, 2005.
7. Levine, D. M. *Statistics for Six Sigma Green Belts*. Upper Saddle River, NJ: Financial Times/Prentice Hall, 2006.
8. Microsoft Excel 2013. Redmond, WA: Microsoft Corp., 2012.
9. Minitab Release 16. State College, PA: Minitab, Inc., 2010.
10. Montgomery, D. M. *Design and Analysis of Experiments*, 6th ed. New York: Wiley, 2005.

KEY EQUATIONS

Total Variation in One-Way ANOVA

$$SST = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{\bar{X}})^2 \quad (11.1)$$

Among-Group Variation in One-Way ANOVA

$$SSA = \sum_{j=1}^c n_j (\bar{X}_j - \bar{\bar{X}})^2 \quad (11.2)$$

Within-Group Variation in One-Way ANOVA

$$SSW = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 \quad (11.3)$$

Mean Squares in One-Way ANOVA

$$MSA = \frac{SSA}{c - 1} \quad (11.4a)$$

$$MSW = \frac{SSW}{n - c} \quad (11.4b)$$

$$MST = \frac{SST}{n - 1} \quad (11.4c)$$

One-Way ANOVA F_{STAT} Test Statistic

$$F_{STAT} = \frac{MSA}{MSW} \quad (11.5)$$

Critical Range for the Tukey-Kramer Procedure

$$\text{Critical range} = Q_\alpha \sqrt{\frac{MSW}{2} \left(\frac{1}{n_j} + \frac{1}{n_{j'}} \right)} \quad (11.6)$$

Total Variation in the Randomized Block Design

$$SST = \sum_{j=1}^c \sum_{i=1}^r (X_{ij} - \bar{\bar{X}})^2 \quad (11.7)$$

Among-Group Variation in the Randomized Block Design

$$SSA = r \sum_{j=1}^c (\bar{X}_{.j} - \bar{\bar{X}})^2 \quad (11.8)$$

Among-Block Variation in the Randomized Block Design

$$SSBL = c \sum_{i=1}^r (\bar{X}_{i.} - \bar{\bar{X}})^2 \quad (11.9)$$

Random Variation in the Randomized Block Design

$$SSE = \sum_{j=1}^c \sum_{i=1}^r (X_{ij} - \bar{X}_j - \bar{X}_{i.} + \bar{\bar{X}})^2 \quad (11.10)$$

Mean Squares in the Randomized Block Design

$$MSA = \frac{SSA}{c - 1} \quad (11.11a)$$

$$MSBL = \frac{SSBL}{r - 1} \quad (11.11b)$$

$$MSE = \frac{SSE}{(r - 1)(c - 1)} \quad (11.11c)$$

F_{STAT} Statistic for Factor Effect

$$F_{STAT} = \frac{MSA}{MSE} \quad (11.12)$$

F_{STAT} Statistic for Block Effects

$$F_{STAT} = \frac{MSBL}{MSE} \quad (11.13)$$

Estimated Relative Efficiency

$$RE = \frac{(r - 1)MSBL + r(c - 1)MSE}{(rc - 1)MSE} \quad (11.14)$$

Critical Range for the Randomized Block Design

$$\text{Critical range} = Q_\alpha \sqrt{\frac{MSE}{r}} \quad (11.15)$$

Total Variation in Two-Way ANOVA

$$SST = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n'} (X_{ijk} - \bar{\bar{X}})^2 \quad (11.16)$$

Factor A Variation

$$SSA = cn' \sum_{i=1}^r (\bar{X}_{i..} - \bar{\bar{X}})^2 \quad (11.17)$$

Factor B Variation

$$SSB = rn' \sum_{j=1}^c (\bar{X}_{.j.} - \bar{\bar{X}})^2 \quad (11.18)$$

Interaction Variation

$$SSAB = n' \sum_{i=1}^r \sum_{j=1}^c (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{\bar{X}})^2 \quad (11.19)$$

Random Variation in Two-Way ANOVA

$$SSE = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n'} (X_{ijk} - \bar{X}_{ij.})^2 \quad (11.20)$$

Mean Squares in Two-Way ANOVA

$$MSA = \frac{SSA}{r - 1} \quad (11.21a)$$

$$MSB = \frac{SSB}{c - 1} \quad (11.21b)$$

$$MSAB = \frac{SSAB}{(r - 1)(c - 1)} \quad (11.21c)$$

$$MSE = \frac{SSE}{rc(n' - 1)} \quad (11.21d)$$

F Test for Factor A Effect

$$F_{STAT} = \frac{MSA}{MSE} \quad (11.22)$$

F Test for Factor B Effect

$$F_{STAT} = \frac{MSB}{MSE} \quad (11.23)$$

F Test for Interaction Effect

$$F_{STAT} = \frac{MSAB}{MSE} \quad (11.24)$$

Critical Range for Factor A

$$\text{Critical range} = Q_\alpha \sqrt{\frac{MSE}{cn'}} \quad (11.25)$$

Critical Range for Factor B

$$\text{Critical range} = Q_\alpha \sqrt{\frac{MSE}{rn'}} \quad (11.26)$$

KEY TERMS

among-block variation 411
 among-group variation 396
 analysis of variance (ANOVA) 395
 ANOVA summary table 399
 blocks 410
 cell means 426
 completely randomized design 395
 critical range 403
 estimated relative efficiency 415

F distribution 398
 factor 395
 grand mean, $\bar{\bar{X}}$ 396
 groups 395
 homogeneity of variance 405
 interaction 419
 levels 395
 Levene test 405
 main effect 423

mean squares 397
 multiple comparisons 402
 normality 405
 one-way ANOVA 395
 randomized block design 410
 randomness and independence 405
 replicates 419
 Studentized range distribution 403
 sum of squares among blocks (SSBL) 411

sum of squares among groups (<i>SSA</i>) 397	sum of squares within groups (<i>SSW</i>) 397	Tukey-Kramer multiple comparisons
sum of squares due to factor <i>A</i> (<i>SSA</i>) 419	total variation 396	procedure for one-way ANOVA 402
sum of squares due to factor <i>B</i> (<i>SSB</i>) 420	Tukey multiple comparison procedure for randomized block designs 415	two-factor factorial design 418
sum of squares due to interaction (<i>SSAB</i>) 420	Tukey multiple comparisons procedure for two-way ANOVA 424	two-way ANOVA 418
sum of squares error (<i>SSE</i>) 411		within-group variation 396
sum of squares total (<i>SST</i>) 395		

CHECKING YOUR UNDERSTANDING

- 11.41** In a one-way ANOVA, what is the difference between the among-groups variance *MSA* and the within-groups variance *MSW*?
- 11.42** What is the difference between the completely randomized one-way ANOVA design and the randomized block design?
- 11.43** What are the distinguishing features of the completely randomized design, the randomized block design, and two-factor factorial designs?
- 11.44** What are the assumptions of ANOVA?
- 11.45** Under what conditions should you use the one-way ANOVA *F* test to examine possible differences among the means of *c* independent populations?
- 11.46** When and how should you use multiple comparison procedures for evaluating pairwise combinations of the group means?
- 11.47** What is the difference between the randomized block design and the two-factor factorial design?
- 11.48** What is the difference between the one-way ANOVA *F* test and the Levene test?
- 11.49** Under what conditions should you use the two-way ANOVA *F* test to examine possible differences among the means of each factor in a factorial design?
- 11.50** What is meant by the concept of interaction in a two-factor factorial design?
- 11.51** How can you determine whether there is an interaction in the two-factor factorial design?

CHAPTER REVIEW PROBLEMS

11.52 You are the production manager at a parachute manufacturing company. Parachutes are woven in your factory using a synthetic fiber purchased from one of four different suppliers. The strength of these fibers is an important characteristic that ensures quality parachutes. You need to decide whether the synthetic fibers from each of your four suppliers result in parachutes of equal strength. Furthermore, to produce parachutes your factory uses two types of looms, the Jetta and the Turk. You need to determine if the parachutes woven on each type of loom are equally strong. You also want to know if any differences in the strength of the parachute can be attributed to the four suppliers are dependent on the type of loom used. You conduct an experiment in which five different parachutes from each supplier are manufactured on each of the two different looms and collect and store the data in [ParachuteTwoWay](#).

At the 0.05 level of significance,

- a. is there an interaction between supplier and loom?
- b. is there an effect due to loom?
- c. is there an effect due to supplier?
- d. Plot the mean strength for each supplier for each loom.
- e. If appropriate, use the Tukey procedure to determine differences between suppliers.
- f. Repeat the analysis, using the [Parachuteoneway](#) file with suppliers as the only factor. Compare your results to those of (b) and (e).

11.53 Medical wires are used in the manufacture of cardiovascular devices. A study was conducted to determine the effect of several factors on the ratio of the load on a test specimen (YS) to the ultimate tensile strength (UTS). The file [MedicalWires1](#) contains the study results, which examined factors including the machine (W95 vs. W96) and the reduction angle (narrow vs. wide). (Data extracted from B. Nepal, S. Mohanty, and L. Kay, "Quality Improvement of Medical Wire Manufacturing Process," *Quality Engineering* 25, 2013, pp. 151–163.)

At the 0.05 level of significance,

- a. is there an interaction between machine type and reduction angle?
- b. is there an effect due to machine type?
- c. is there an effect due to reduction angle?
- d. Plot the mean ratio of the load on a test specimen (YS) to the ultimate tensile strength (UTS) for each machine type for each reduction angle.
- e. What can you conclude about the effects of machine type and reduction angle on the ratio of the load on a test specimen (YS) to the ultimate tensile strength (UTS)? Explain.
- f. Repeat the analysis, using reduction angle as the only factor (see the [MedicalWires2](#) file. Compare your results to those of (c) and (e).

11.54 An operations manager wants to examine the effect of air-jet pressure (in pounds per square inch [psi]) on the breaking strength of yarn. Three different levels of air-jet pressure are to be considered: 30 psi, 40 psi, and 50 psi. A random sample of 18 yarns are selected from the same batch, and the yarns are randomly assigned, 6 each, to the 3 levels of air-jet pressure. The breaking strength scores are stored in **Yarn**.

- Is there evidence of a significant difference in the variances of the breaking strengths for the three air-jet pressures? (Use $\alpha = 0.05$.)
- At the 0.05 level of significance, is there evidence of a difference among mean breaking strengths for the three air-jet pressures?
- If appropriate, use the Tukey-Kramer procedure to determine which air-jet pressures significantly differ with respect to mean breaking strength. (Use $\alpha = 0.05$.)
- What should the operations manager conclude?

11.55 Suppose that, when setting up the experiment in Problem 11.54, the operations manager is able to study the effect of side-to-side aspect in addition to air-jet pressure. Thus, instead of the one-factor completely randomized design in Problem 11.54, a two-factor factorial design was used, with the first factor, side-to-side aspect, having two levels (nozzle and opposite) and the second factor, air-jet pressure, having three levels (30 psi, 40 psi, and 50 psi). A sample of 18 yarns is randomly assigned, 3 to each of the 6 side-to-side aspect and pressure level combinations. The breaking-strength scores, stored in **Yarn**, are as follows:

SIDE-TO-SIDE ASPECT	AIR-JET PRESSURE		
	30 psi	40 psi	50 psi
Nozzle	25.5	24.8	23.2
Nozzle	24.9	23.7	23.7
Nozzle	26.1	24.4	22.7
Opposite	24.7	23.6	22.6
Opposite	24.2	23.3	22.8
Opposite	23.6	21.4	24.9

At the 0.05 level of significance,

- is there an interaction between side-to-side aspect and air-jet pressure?
- is there an effect due to side-to-side aspect?
- is there an effect due to air-jet pressure?
- Plot the mean yarn breaking strength for each level of side-to-side aspect for each level of air-jet pressure.
- If appropriate, use the Tukey procedure to study differences among the air-jet pressures.
- On the basis of the results of (a) through (e), what conclusions can you reach concerning yarn breaking strength? Discuss.
- Compare your results in (a) through (f) with those from the completely randomized design in Problem 11.38. Discuss fully.

11.56 A hotel wanted to develop a new system for delivering room service breakfasts. In the current system, an order form is left on the

bed in each room. If the customer wishes to receive a room service breakfast, he or she places the order form on the doorknob before 11 P.M. The current system requires customers to select a 15-minute interval for desired delivery time (6:30–6:45 A.M., 6:45–7:00 A.M., etc.). The new system is designed to allow the customer to request a specific delivery time. The hotel wants to measure the difference (in minutes) between the actual delivery time and the requested delivery time of room service orders for breakfast. (A negative time means that the order was delivered before the requested time. A positive time means that the order was delivered after the requested time.) The factors included were the menu choice (American or Continental) and the desired time period in which the order was to be delivered (Early Time Period [6:30–8:00 A.M.] or Late Time Period [8:00–9:30 A.M.]). Ten orders for each combination of menu choice and desired time period were studied on a particular day. The data, stored in **Breakfast**, are as follows:

TYPE OF BREAKFAST	DESIRED TIME	
	Early Time Period	Late Time Period
Continental	1.2	-2.5
Continental	2.1	3.0
Continental	3.3	-0.2
Continental	4.4	1.2
Continental	3.4	1.2
Continental	5.3	0.7
Continental	2.2	-1.3
Continental	1.0	0.2
Continental	5.4	-0.5
Continental	1.4	3.8
American	4.4	6.0
American	1.1	2.3
American	4.8	4.2
American	7.1	3.8
American	6.7	5.5
American	5.6	1.8
American	9.5	5.1
American	4.1	4.2
American	7.9	4.9
American	9.4	4.0

At the 0.05 level of significance,

- is there an interaction between type of breakfast and desired time?
- is there an effect due to type of breakfast?
- is there an effect due to desired time?
- Plot the mean delivery time difference for each desired time for each type of breakfast.
- On the basis of the results of (a) through (d), what conclusions can you reach concerning delivery time difference? Discuss.

11.57 Refer to the room service experiment in Problem 11.56. Now suppose that the results are as shown below and stored in **Breakfast2**. Repeat (a) through (e), using these data, and compare the results to those of (a) through (e) of Problem 11.56.

TYPE OF BREAKFAST	DESIRED TIME	
	Early	Late
Continental	1.2	-0.5
Continental	2.1	5.0
Continental	3.3	1.8
Continental	4.4	3.2
Continental	3.4	3.2
Continental	5.3	2.7
Continental	2.2	0.7
Continental	1.0	2.2
Continental	5.4	1.5
Continental	1.4	5.8
American	4.4	6.0
American	1.1	2.3
American	4.8	4.2
American	7.1	3.8
American	6.7	5.5
American	5.6	1.8
American	9.5	5.1
American	4.1	4.2
American	7.9	4.9
American	9.4	4.0

11.58 A pet food company has the business objective of having the weight of a can of cat food come as close to the specified weight as possible. Realizing that the size of the pieces of meat contained in a can and the can fill height could impact the weight of a can, a team studying the weight of canned cat food wondered whether the current larger chunk size produced higher can weight and more variability. The team decided to study the effect on weight of a cutting size that was finer than the current size. In addition, the team slightly lowered the target for the sensing mechanism that determines the fill height in order to determine the effect of the fill height on can weight.

Twenty cans were filled for each of the four combinations of piece size (fine and current) and fill height (low and current). The contents of each can were weighed, and the amount above or below the label weight of 3 ounces was recorded as the variable

coded weight. For example, a can containing 2.90 ounces was given a coded weight of -0.10. Results were stored in **CatFood2**.

Analyze these data and write a report for presentation to the team. Indicate the importance of the piece size and the fill height on the weight of the canned cat food. Be sure to include a recommendation for the level of each factor that will come closest to meeting the target weight and the limitations of this experiment, along with recommendations for future experiments that might be undertaken.

11.59 Brand valuations are critical to CEOs, financial and marketing executives, security analysts, institutional investors, and others who depend on well-researched, reliable information needed for assessments and comparisons in decision making. Millward Brown Optimor has developed the BrandZ Top 100 Most Valuable Global Brands for WPP, the world's largest communications services group. Unlike other studies, the BrandZ Top 100 Most Valuable Global Brands fuses consumer measures of brand equity with financial measures to place a financial value on brands. The file **BrandZTechFinTele** contains the brand values for three sectors in the BrandZ Top 100 Most Valuable Global Brands for 2013: the technology sector, the financial institutions sector, and the telecom sector. (Data extracted from bit.ly/18OL5Mu.)

- At the 0.01 level of significance, is there evidence of a difference in mean brand value among the sectors?
- What assumptions are necessary in order to complete (a)? Comment on the validity of these assumptions.
- If appropriate, use the Tukey procedure to determine the sectors that differ in mean rating. (Use $\alpha = 0.01$.)

11.60 An investor can choose from a very large number of mutual funds. Each mutual fund has its own mix of different types of investments. The data in **BestFunds2** present the one-year return and the three-year annualized return for the 10 best short-term bond, long-term bond, and world bond funds, according to the *U.S. News & World Report*. (Data extracted from money.usnews.com/mutual-funds/rankings.) Analyze the data and determine whether any differences exist in the one-year return and the three-year annualized return between short-term, long-term, and world bond funds. (Use the 0.05 level of significance.)

11.61 An investor can choose from a very large number of mutual funds. Each mutual fund has its own mix of different types of investments. The data in **BestFunds3** present the one-year return and the three-year annualized return for the 10 best small cap growth, mid-cap growth, and large cap growth funds, according to the *U.S. News & World Report*. (Data extracted from money.usnews.com/mutual-funds/rankings.) Analyze the data and determine whether any differences exist in the one-year return and the three-year annualized return between small cap growth, mid-cap growth, and large cap growth funds. (Use the 0.05 level of significance.)

CASES FOR CHAPTER 11

Managing Ashland MultiComm Services

PHASE 1

The computer operations department had a business objective of reducing the amount of time to fully update each subscriber's set of messages in a special secured email system. An experiment was conducted in which 24 subscribers were selected and three different messaging systems were used. Eight subscribers were assigned to each system, and the update times were measured. The results, stored in **AMS11-1**, are presented in Table AMS11.1.

TABLE AMS11.1

Update Times (in seconds) for Three Different Systems

System 1	System 2	System 3
38.8	41.8	32.9
42.1	36.4	36.1
45.2	39.1	39.2
34.8	28.7	29.3
48.3	36.4	41.9
37.8	36.1	31.7
41.1	35.8	35.2
43.6	33.7	38.1

- Analyze the data in Table AMS11.1 and write a report to the computer operations department that indicates your findings. Include an appendix in which you discuss the reason you selected a particular statistical test to compare the three email interfaces.

DO NOT CONTINUE UNTIL THE PHASE 1 EXERCISE HAS BEEN COMPLETED.

PHASE 2

After analyzing the data in Table AMS11.1, the computer operations department team decided to also study the effect of the connection media used (cable or fiber).

The team designed a study in which a total of 30 subscribers were chosen. The subscribers were randomly assigned to one of the three messaging systems so that there were five subscribers in each of the six combinations of the two factors—messaging system and media used. Measurements were taken on the updated time. Table AMS11.2 summarizes the results that are stored in **AMS11-2**.

TABLE AMS11.2

Update Times (in seconds), Based on Messaging System and Media Used

MEDIA	INTERFACE		
	System 1	System 2	System 3
Cable	4.56	4.17	3.53
	4.90	4.28	3.77
	4.18	4.00	4.10
	3.56	3.96	2.87
	4.34	3.60	3.18
	4.41	3.79	4.33
Fiber	4.08	4.11	4.00
	4.69	3.58	4.31
	5.18	4.53	3.96
	4.85	4.02	3.32

- Completely analyze these data and write a report to the team that indicates the importance of each of the two factors and/or the interaction between them on the update time. Include recommendations for future experiments to perform.

Digital Case

Apply your knowledge about ANOVA in this Digital Case, which continues the cereal-fill packaging dispute Digital Case from Chapters 7, 9, and 10.

After reviewing CCACC's latest document (see the Digital Case for Chapter 10 on page 384), Oxford Cereals has released **SecondAnalysis.pdf**, a press kit that Oxford Cereals has assembled to refute the claim that it is guilty of using selective data. Review the Oxford Cereals press kit and then answer the following questions.

- Does Oxford Cereals have a legitimate argument? Why or why not?
- Assuming that the samples Oxford Cereals has posted were randomly selected, perform the appropriate analysis to resolve the ongoing weight dispute.
- What conclusions can you reach from your results? If you were called as an expert witness, would you support the claims of the CCACC or the claims of Oxford Cereals? Explain.

Sure Value Convenience Stores

You work in the corporate office for a nationwide convenience store franchise that operates nearly 10,000 stores. The per-store daily customer count (i.e., the mean number of customers in a store in one day) has been steady, at 900, for some time. To increase the customer count, the chain is considering cutting prices for coffee beverages. The question to be determined is how much to cut prices to increase the daily customer count without reducing the gross margin on coffee sales too much. You decide to carry out an experiment in a sample of 24 stores where customer counts have been running almost exactly at the national average of 900. In 6 of the stores, the price of a small coffee will now be \$0.59, in 6 stores the price of a small coffee will now be

\$0.69, in 6 stores, the price of a small coffee will now be \$0.79, and in 6 stores, the price of a small coffee will now be \$0.89. After four weeks of selling the coffee at the new price, the daily customer counts in the stores were recorded and stored in **CoffeeSales**.

- Analyze the data and determine whether there is evidence of a difference in the daily customer count, based on the price of a small coffee.
- If appropriate, determine which mean prices differ in daily customer counts.
- What price do you recommend for a small coffee?

CardioGood Fitness

Return to the CardioGood Fitness case (stored in **CardioGood Fitness**) first presented on page 83.

- Determine whether differences exist between customers based on the product purchased (TM195, TM498, TM798) in their age in years, education in years, annual household income (\$), mean number of times the

customer plans to use the treadmill each week, and mean number of miles the customer expects to walk or run each week.

- Write a report to be presented to the management of CardioGood Fitness detailing your findings.

More Descriptive Choices Follow-Up

Follow up the Using Statistics scenario “More Descriptive Choices, Revisited” on page 138 by determining whether there is a difference between the small, mid-cap, and large

market cap funds in the three-year return percentages, five-year return percentages, and ten-year return percentages (stored in **Retirement Funds**).

Clear Mountain State Student Surveys

- The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students who attend CMSU. They create and distribute a survey of 14 questions and receive responses from 62 undergraduates (stored in **UndergradSurvey**).
 - At the 0.05 level of significance, is there evidence of a difference based on academic major in expected starting salary, number of social networking sites reg-

istered for, age, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?

- At the 0.05 level of significance, is there evidence of a difference based on graduate school intention in grade point average, expected starting salary, number of social networking sites registered for, age, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?

2. The dean of students at CMSU has learned about the undergraduate survey and has decided to undertake a similar survey for graduate students at Clear Mountain State. She creates and distributes a survey of 14 questions and receives responses from 44 graduate students (stored in **GradSurvey**). For these data, at the 0.05 level of significance,
- a. is there evidence of a difference based on undergraduate major in age, undergraduate grade point average, graduate grade point average, expected salary upon graduation, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?
 - b. is there evidence of a difference based on graduate major in age, undergraduate grade point average, graduate grade point average, expected salary upon graduation, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?
 - c. is there evidence of a difference based on employment status in age, undergraduate grade point average, graduate grade point average, expected salary upon graduation, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?

CHAPTER 11 EXCEL GUIDE

EG11.1 The COMPLETELY RANDOMIZED DESIGN: ONE-WAY ANOVA

Analyzing Variation in One-Way ANOVA

Key Technique Use the Section EG2.5 instructions to construct scatter plots using stacked data. If necessary, change the levels of the factor to consecutive integers beginning with 1, as was done for the in-store location sales experiment data in Figure 11.4 on page 400.

F Test for Differences Among More Than Two Means

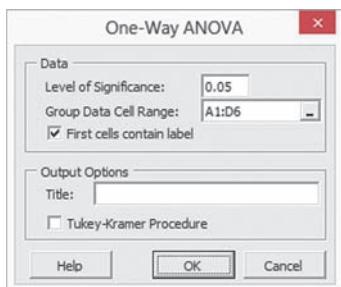
Key Technique Use the **DEVSQ** (*cell range of data of all groups*) function to compute *SST* and uses an expression in the form *SST – DEVSQ (group 1 data cell range) – DEVSQ (group 2 data cell range) ... – DEVSQ (group n data cell range)* to compute *SSA*.

Example Perform the Figure 11.6 one-way ANOVA for the in-store location sales experiment shown on page 402.

PHStat Use One-Way ANOVA.

For the example, open to the **DATA worksheet** of the **Mobile Electronics workbook**. Select **PHStat → Multiple-Sample Tests → One-Way ANOVA**. In the procedure's dialog box (shown below):

1. Enter **0.05** as the **Level of Significance**.
2. Enter **A1:D6** as the **Group Data Cell Range**.
3. Check **First cells contain label**.
4. Enter a **Title**, clear the **Tukey-Kramer Procedure** check box, and click **OK**.



In addition to the worksheet shown in Figure 11.6, this procedure creates an **ASFData worksheet** to hold the data used for the test. See the following *In-Depth Excel* section for a complete description of this worksheet.

In-Depth Excel Use the COMPUTE worksheet of the One-Way ANOVA workbook as a template.

The COMPUTE worksheet, and the supporting ASFData worksheet, already contains the data for the example. Modifying the One-Way ANOVA workbook for use with other problems is

more difficult than modifications discussed in the previous Excel Guides. To modify the workbook:

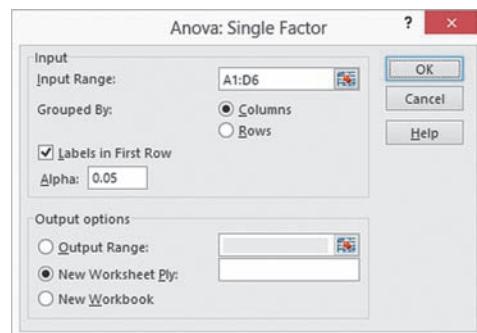
1. Paste the data for the problem into the **ASFData worksheet**, overwriting the in-store locations sales experiment data.
- In the COMPUTE worksheet (see Figure 11.6):
 2. Edit the *SST* formula = **DEVSQ(ASFData!A1:D6)** in cell B16 to use the cell range of the new data just pasted into the ASFData worksheet.
 3. Edit the cell B13 *SSA* formula so there are as many **DEVSQ(group column cell range)** terms as there are groups.
 4. Change the level of significance in cell G17, if necessary.
 5. If the problem contains three groups, select **row 8**, right-click, and select **Delete** from the shortcut menu. If the problem contains more than four groups, select **row 8**, right-click, and click **Insert** from the shortcut menu. Repeat this step as many times as necessary.
 6. If you inserted new rows, enter (not copy) the formulas for those rows, using the formulas in row 7 as models.
 7. Adjust table formatting as necessary.

Read the **SHORT TAKES** for Chapter 11 for an explanation of the formulas found in the COMPUTE worksheet (shown in the **COMPUTE_FORMULAS worksheet**). If you use an Excel version older than Excel 2010, use the **COMPUTE_OLEDER** worksheet.

Analysis ToolPak Use Anova: Single Factor.

For the example, open to the **DATA worksheet** of the **Mobile Electronics workbook** and:

1. Select **Data → Data Analysis**.
2. In the Data Analysis dialog box, select **Anova: Single Factor** from the **Analysis Tools** list and then click **OK**.
- In the procedure's dialog box (shown below):
 3. Enter **A1:D6** as the **Input Range**.
 4. Click **Columns**, check **Labels in First Row**, and enter **0.05** as **Alpha**.
 5. Click **New Worksheet Ply**.
 6. Click **OK**.



The Analysis ToolPak creates a worksheet that does not use formulas but is similar in layout to the Figure 11.6 worksheet on page 402.

Multiple Comparisons: The Tukey-Kramer Procedure

Key Technique Use formulas to compute the absolute mean differences and use the **IF** function to compare pairs of means.

Example Perform the Figure 11.7 Tukey-Kramer procedure for the in-store location sales experiment shown on page 402.

PHStat Use the *PHStat* instructions for the one-way ANOVA *F* test to perform the Tukey-Kramer procedure, *checking Tukey-Kramer Procedure* instead in step 4. The procedure creates a worksheet identical to the one shown in Figure 11.7 on page 404 and discussed in the following *In-Depth Excel* section. To complete the worksheet, enter the Studentized range *Q statistic* (use Table E.7) for the level of significance and the numerator and denominator degrees of freedom that are given in the worksheet.

In-Depth Excel To perform the Tukey-Kramer procedure, first use the *In-Depth Excel* instructions for the one-way ANOVA *F* test and then use the appropriate “**TK**” **worksheet** in the **One-Way ANOVA workbook**.

For the example, open to the **TK4 worksheet** that already has the value of the *Q* statistic (4.05) entered in cell B15.

The TK worksheets can be used for problems using three (**TK3**), four (**TK4**), five (**TK5**), six (**TK6**), or seven (**TK7**) groups. Use Table E.7 to look up the proper value of the Studentized range *Q* statistic for the level of significance and the numerator and denominator degrees of freedom for the problem. When you use either the **TK5**, **TK6**, and **TK7** worksheets, you must also enter the name, sample mean, and sample size for the fifth and, if applicable, sixth and seventh groups.

Read the **SHORT TAKES** for Chapter 11 for an explanation of the formulas found in the COMPUTE worksheet (shown in the **COMPUTE_FORMULAS worksheet**). If you use an Excel version older than Excel 2010, use the **TK4_OLDER** worksheet.

Analysis ToolPak Modify the previous *In-Depth Excel* instructions to perform the Tukey-Kramer procedure in conjunction with using the **Anova: Single Factor** procedure. Transfer selected values from the Analysis ToolPak results worksheet to one of the TK worksheets in the **One-Way ANOVA workbook**. For example, to perform the Figure 11.7 Tukey-Kramer procedure for the in-store location sales experiment on page 404:

1. Use the **Anova: Single Factor** procedure, as described earlier in this section, to create a worksheet that contains ANOVA results for the in-store locations experiment.
2. Record the name, **sample size** (in the **Count** column), and **sample mean** (in the **Average** column) of each group. Also record the **MSW** value, found in the cell that is the intersection of the **MS** column and **Within Groups** row, and the **denominator degrees of freedom**, found in the cell that is the intersection of the **df** column and **Within Groups** row.

degrees of freedom, found in the cell that is the intersection of the **df** column and **Within Groups** row.

3. Open to the **TK4 worksheet** of the **One-Way ANOVA workbook**.

In the **TK4** worksheet:

4. Overwrite the formulas in cell range A5:C8 by entering the name, sample mean, and sample size of each group into that range.
5. Enter **0.05** as the **Level of significance** in cell B11.
6. Enter **4** as the **Numerator d.f.** (equal to the number of groups) in cell B12.
7. Enter **16** as the **Denominator d.f.** in cell B13.
8. Enter **0.0439** as the **MSW** in cell B14.
9. Enter **4.05** as the **Q Statistic** in cell B15. (Look up the Studentized range *Q* statistic using Table E.7.)

Levene Test for Homogeneity of Variance

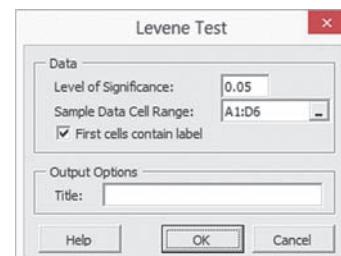
Key Technique Use the techniques for performing a one-way ANOVA.

Example Perform the Figure 11.8 Levene test for the in-store location sales experiment shown on page 406.

PHStat Use **Levene Test**.

For the example, open to the **DATA worksheet** of the **Mobile Electronics workbook**. Select **PHStat → Multiple-Sample Tests → Levene Test**. In the procedure’s dialog box (shown below):

1. Enter **0.05** as the **Level of Significance**.
2. Enter **A1:D6** as the **Sample Data Cell Range**.
3. Check **First cells contain label**.
4. Enter a **Title** and click **OK**.



The procedure creates a worksheet that performs the Table 11.4 absolute differences computations (see page 406) as well as the Figure 11.8 worksheet. See the following *In-Depth Excel* section for a description of these worksheets.

In-Depth Excel Use the **COMPUTE worksheet** of the **Levene workbook** as a template.

The COMPUTE worksheet and the supporting AbsDiffs and DATA worksheets already contain the data for the example.

For other problems in which the absolute differences are already known, paste the absolute differences into the AbsDiffs

worksheet. Otherwise, paste the problem data into the DATA worksheet, add formulas to compute the median for each group, and adjust the AbsDiffs worksheet as necessary. For example, for the in-store location sales experiment, the following steps 1 through 7 were done with the workbook open to the DATA worksheet:

1. Enter the label **Medians** in cell A7, the first empty cell in column A.
2. Enter the formula **=MEDIAN(A2:A6)** in cell A8. (Cell range A2:A6 contains the data for the first group, Supplier 1.)
3. Copy the cell A8 formula across through column D.
4. Open to the **AbsDiffs** worksheet.

In the AbsDiffs worksheet:

5. Enter row 1 column headings **AbsDiff1**, **AbsDiff2**, **AbsDiff3**, and **AbsDiff4** in columns A through D.
6. Enter the formula **=ABS(DATA!A2 - DATA!A8)** in cell A2. Copy this formula down through row 6.
7. Copy the formulas now in cell range A2:A6 across through column D. Absolute differences now appear in the cell range A2:D6.

If you use an Excel version older than Excel 2010, use the COMPUTE_OLDER worksheet.

Analysis ToolPak Use **Anova: Single Factor** with absolute difference data to perform the Levene test. If the absolute differences have not already been computed, use steps 1 through 7 of the preceding *In-Depth Excel* instructions to compute them.

EG11.2 The RANDOMIZED BLOCK DESIGN

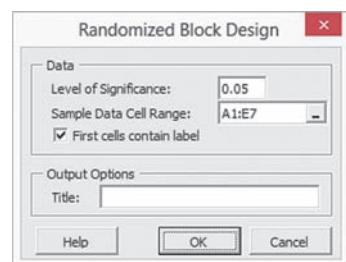
Key Technique Use the **F.INV.RT**, **E.DIST.RT**, and **DEVSQ** functions to help compute the ANOVA summary table statistics for a randomized block design. Enter **F.INV.RT**(*level of significance, degrees of freedom for source, Error degrees of freedom*) to compute the F critical value for the among-groups (*A*) and among-blocks (*BL*) sources of variation. Enter **E.DIST.RT**(*F test statistic for source, degrees of freedom for rows, Error degrees of freedom within groups*) to calculate the *p*-value for the two sources of variation. Use the **DEVSQ** function to compute *SSA*, *SSBL*, *SSE*, and *SST*.

Example Perform the Figure 11.10 randomized block design for the quick-service restaurant chain study on page 414.

PHStat Use **Randomized Block Design**.

For the example, open to the **DATA worksheet** of the **QSR Chain** **workbook**. Select **PHStat** → **Multiple-Sample Tests** → **Randomized Block Design**. In the procedure's dialog box (shown at top right):

1. Enter **0.05** as the **Level of Significance**.
2. Enter **A1:E7** as the **Sample Data Cell Range**.
3. Check **First cells contain label**.
4. Enter a **Title** and click **OK**.



This procedure requires that the labels that identify factor *A* appear stacked in column A, followed by columns for factor *B*.

In-Depth Excel Use the **COMPUTE worksheet** of the **Randomized Block workbook** as a model.

For the example, the worksheet already uses the contents of the DATA worksheet to perform the test for the example. In the worksheet ANOVA summary table, the source labeled **Among groups (A)** in Table 11.6 on page 413 is labeled **Columns**, and the source **Among blocks (BL)** is labeled **Rows**.

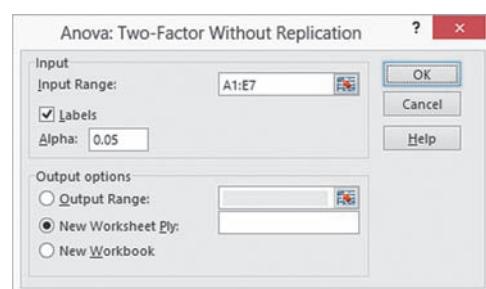
For problems with the same number of levels for each factor as the example, paste the data for the problem into the **DATA worksheet**, overwriting the restaurant rating data. Because of the complexity of the COMPUTE worksheet, consider using either PHStat or the Analysis ToolPak for other problems that have a different mix of factors and levels. If that is not possible, use the instructions in the **SHORT TAKES** for Chapter 11 to modify the Randomized Block workbook.

Read the **SHORT TAKES** for Chapter 11 for an explanation of the formulas found in the COMPUTE worksheet (shown in the **COMPUTE_FORMULAS worksheet**). If you are using an older Excel version, use the COMPUTE_OLDER worksheet instead of the COMPUTE worksheet.

Analysis ToolPak Use the **Anova: Two-Factor Without Replication**.

For the example, open to the **DATA worksheet** of the **QSRChain** **workbook** and:

1. Select **Data** → **Data Analysis**.
 2. In the Data Analysis dialog box, select **Anova: Two-Factor Without Replication** from the **Analysis Tools** list and then click **OK**.
- In the procedure's dialog box (shown below):
3. Enter **A1:E7** as the **Input Range**.
 4. Check **Labels** and enter **0.05** as **Alpha**.
 5. Click **New Worksheet Ply**.
 6. Click **OK** to create the worksheet.



This procedure requires that the labels that identify blocks appear stacked in column A and that group names appear in row 1, starting with cell B1.

The Analysis ToolPak creates a worksheet that is visually similar to Figure 11.10 but contains only values and does not include any cell formulas. The ToolPak worksheet also does not contain the level of significance in row 24.

EG11.3 The FACTORIAL DESIGN: TWO-WAY ANOVA

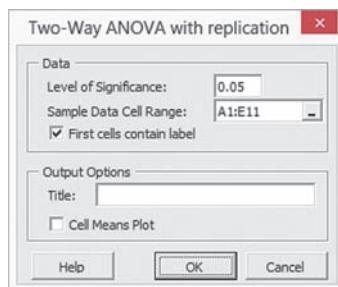
Key Technique Use the **DEVSQ** function to compute SSA, SSB, SSAB, SSE, and SST.

Example Perform the Figure 11.14 two-way ANOVA for the in-store location sales and mobile payment experiment shown on page 423.

PHStat Use Two-Way ANOVA with replication.

For the example, open to the **DATA worksheet** of the **Mobile Electronics2 workbook**. Select **PHStat → Multiple-Sample Tests → Two-Way ANOVA**. In the procedure's dialog box (shown below):

1. Enter **0.05** as the **Level of Significance**.
2. Enter **A1:E11** as the **Sample Data Cell Range**.
3. Check **First cells contain label**.
4. Enter a **Title** and click **OK**.



This procedure requires that the labels that identify factor A appear stacked in column A, followed by columns for factor B.

In-Depth Excel Use the **COMPUTE worksheet** of the **Two-Way ANOVA workbook** as a model.

For the example, the worksheet already uses the contents of the **DATA worksheet** to perform the test for the example.

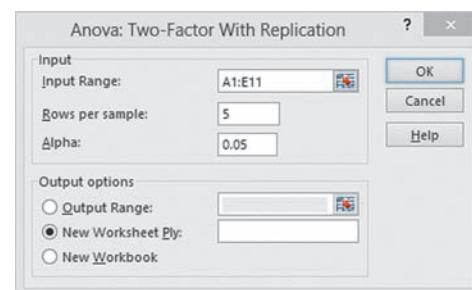
Because of the complexity of the **COMPUTE worksheet**, consider using either PHStat or the Analysis ToolPak for other problems, especially ones that have a different mix of factors and levels. If that is not possible, use the instructions in the **SHORT TAKES** for Chapter 11 to modify the Two-Way workbook. For problems in which $r = 2$ and $c = 4$, paste the data for the problem into the **ATFData worksheet**, overwriting the in-store location and mobile payments data and then adjust the factor level headings in the **COMPUTE worksheet**.

Read the **SHORT TAKES** for Chapter 11 for an explanation of the formulas found in the **COMPUTE worksheet** (shown in the **COMPUTE_FORMULAS worksheet**). If you use an Excel

version older than Excel 2010, use the **COMPUTE_OLDER** worksheet instead of the **COMPUTE** worksheet.

Analysis ToolPak Use **Anova: Two-Factor With Replication**. For the example, open to the **DATA worksheet** of the **Mobile Electronics2 workbook** and:

1. Select **Data → Data Analysis**.
 2. In the Data Analysis dialog box, select **Anova: Two-Factor With Replication** from the **Analysis Tools** list and then click **OK**.
- In the procedure's dialog box (shown below):
3. Enter **A1:E11** as the **Input Range**.
 4. Enter **5** as the **Rows per sample**.
 5. Enter **0.05** as **Alpha**.
 6. Click **New Worksheet Ply**.
 7. Click **OK**.



This procedure requires that the labels that identify factor A appear stacked in column A, followed by columns for factor B. The Analysis ToolPak creates a worksheet that does not use formulas but is similar in layout to the Figure 11.14 worksheet.

Visualizing Interaction Effects: The Cell Means Plot

Key Technique Use the **SUMPRODUCT(cell range 1, cell range 2)** function to compute the expected value and variance.

Example Construct the Figure 11.17 cell means plot for mobile electronics sales based on mobile payments permitted and in-store location on page 426.

PHStat Modify the **PHStat** instructions for the two-way ANOVA. In step 4, check **Cell Means Plot** before clicking **OK**.

In-Depth Excel Create a cell means plot from a two-way ANOVA **COMPUTE** worksheet.

For the example, open to the **COMPUTE worksheet** of the **Two-Way ANOVA** workbook and:

1. Insert a new worksheet.
2. Copy cell range **B3:E3** of the **COMPUTE** worksheet (the factor B level names) to cell **B1** of the new worksheet, using the **Paste Special Values** option.

3. Copy the cell range **B7:E7** of the COMPUTE worksheet (the AVERAGE row for the factor A No level) and paste to cell **B2** of the new worksheet, using the Paste Special **Values** option.
4. Copy the cell range **B13:E13** of the COMPUTE worksheet (the AVERAGE row for the factor A Yes level) and paste to cell **B3** of a new worksheet, using the Paste Special **Values** option.
5. Enter **No** in cell **B3** and **Yes** in cell **A3** of the new worksheet as labels for the factor *A* levels.
6. Select the cell range **A1:E3**.

7. Select **Insert → Line** and select the **fourth 2-D Line gallery choice (Line with Markers)**.

8. Relocate the chart to a chart sheet, add axis titles, and modify the chart title by using the instructions in Appendix Section B.6.

For other problems, insert a new worksheet and first copy and paste the factor *B* level names to row 1 of the new worksheet and then copy and use Paste Special to transfer the values in the **Average** rows data for each factor *B* level to the new worksheet. (See Appendix B to learn more about the Paste Special command.)

Analysis ToolPak Use the *In-Depth Excel* instructions.

CHAPTER 11 MINITAB GUIDE

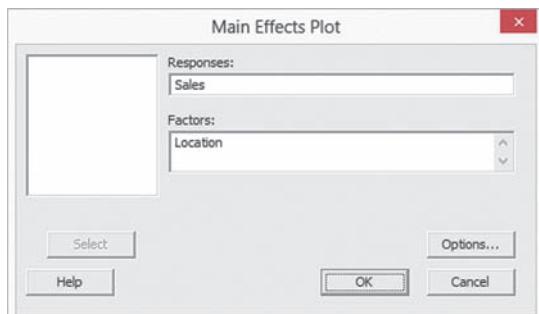
MG11.1 The COMPLETELY RANDOMIZED DESIGN: ONE-WAY ANOVA

Analyzing Variation in One-Way ANOVA

Use **Main Effects Plot** (requires stacked data).

For example, to construct the Figure 11.4 main effects plot for the in-store location sales experiment on page 400, open to the **Mobile Electronics Stacked** worksheet. Select **Stat → ANOVA → Main Effects Plot**. In the Main Effects Plot dialog box (shown below):

1. Double-click **C2 Sales** in the variables list to add **Sales** to the **Responses** box and press **Tab**.
2. Double-click **C1 Location** in the variables list to add **Location** to the **Factors** box.
3. Click **OK**.



In step 2, if the column entered in the Factors box contains a text variable, as it does in the example, Minitab will sort the factor levels alphabetically. To present levels in a different order, as was done in Figure 11.4, right-click one of the factor levels in the chart and click **Edit X Scale** from the shortcut menu. In the Edit Scale dialog box, click **Specified**, type the factor levels in the desired order separated by spaces, and click **OK**.

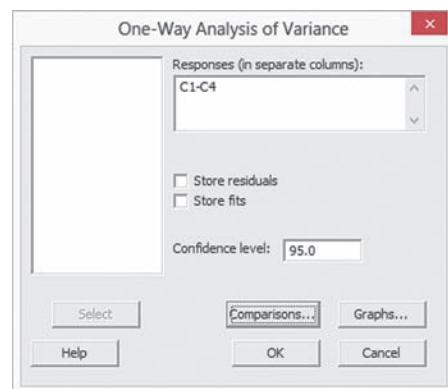
F Test for Differences Among More Than Two Means

Use **One-Way (Unstacked)** or **One-Way** (for stacked data.)

For example, to perform the Figure 11.6 one-way ANOVA for the in-store location sales experiment on page 402, open to the

Mobile Electronics worksheet. Select **Stat → ANOVA → One-Way (Unstacked)**. In the One-Way Analysis of Variance dialog box (shown below):

1. Enter **C1-C4** in the **Responses (in separate columns)** box.
2. Enter **95.0** in the **Confidence level** box.
3. Click **Comparisons**.



In the One-Way Multiple Comparisons dialog box (shown below):

4. Clear all check boxes.
5. Click **OK**.



6. Back in the original dialog box, click **Graphs**.

In the One-Way Analysis of Variance - Graphs dialog box (not shown):

7. Check **Boxplots of data**.
8. Click **OK**.
9. Back in the original dialog box, click **OK**.

When using stacked data, select **Stat** → **ANOVA** → **One-Way** and in step 1 enter the name of the column that contains the variable of interest in the **Response** box and enter the name of the column that contains the factor names in the **Factor** box.

Multiple Comparisons: The Tukey-Kramer Procedure

Use the previous set of instructions to perform the Tukey-Kramer procedure, replacing step 4 with:

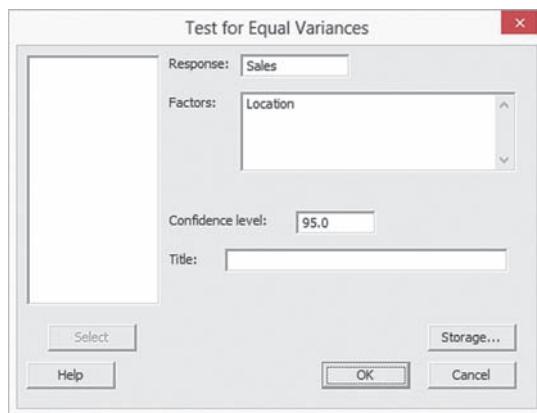
4. Check **Tukey's, family error rate** and enter **5** in its box. (A family error rate of 5 produces comparisons with an overall confidence level of 95%).

Levene Test for Homogeneity of Variance

Use **Test for Equal Variances** (requires stacked data).

For example, to perform the Figure 11.8 Levene test for the in-store location sales experiment on page 406, open to the **Mobile Electronics Stacked** worksheet, which contains the data of the Parachute worksheet in stacked order. Select **Stat** → **ANOVA** → **Test for Equal Variances**. In the Test for Equal Variances dialog box (shown below):

1. Double-click **C2 Sales** in the variables list to add **Sales** to the **Response** box.
2. Double-click **C1 Location** in the variables list to add **Location** to the **Factor** box.
3. Enter **95.0** in the **Confidence level** box.
4. Click **OK**.



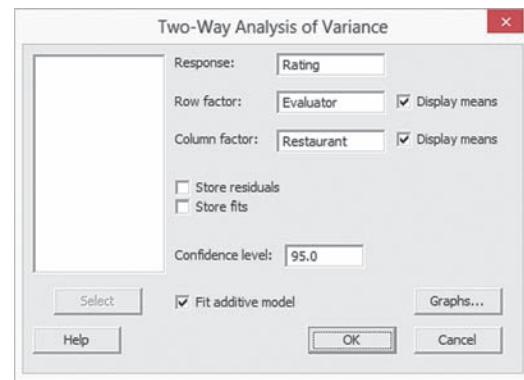
The Levene test results shown in Figure 11.8 on page 406 appear last in the results this procedure creates.

MG11.2 The RANDOMIZED BLOCK DESIGN

Use **Two-Way** (requires stacked data).

For example, to perform the Figure 11.10 randomized block design for the quick-service restaurant chain study shown on page 414, open to the **QSRChain worksheet**. Select **Stat** → **ANOVA** → **Two-Way**. In the Two-Way Analysis of Variance dialog box (shown below):

1. Double-click **C3 Rating** in the variables list to add **Rating** to the **Response** box.
2. Double-click **C1 Evaluator** in the variables list to add **Evaluator** to the **Row factor** box.
3. Double-click **C2 Restaurant** in the variables list to add **Restaurant** to the **Column factor** box.
4. Check **Display means** for both the row and column factors.
5. Enter **95.0** in the **Confidence level** box.
6. Check **Fit Additive Model**.
7. Click **OK**.

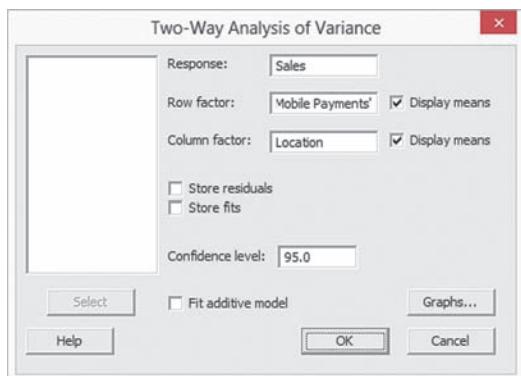


MG11.3 The FACTORIAL DESIGN: TWO-WAY ANOVA

Use **Two-Way** (requires stacked data).

For example, to perform the Figure 11.14 two-way ANOVA for the in-store location sales and mobile payment experiment on page 423, open to the **Mobile Electronics2 worksheet**. Select **Stat** → **ANOVA** → **Two-Way**. In the Two-Way Analysis of Variance dialog box (shown on page 446):

1. Double-click **C3 Sales** in the variables list to add **Sales** to the **Response** box.
2. Double-click **C1 Mobile Payments** in the variables list to add **Mobile Payments** to the **Row factor** box.
3. Double-click **C2 Location** in the variables list to add **Location** to the **Column factor** box.
4. Check **Display Means** for both the row and column factors.
5. Enter **95.0** in the **Confidence level** box.
6. Click **OK**.

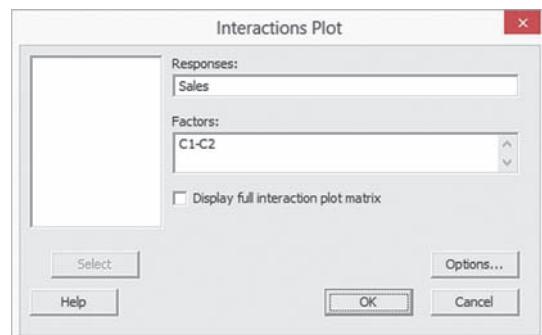


Visualizing Interaction Effects: The Cell Means Plot

Use **Interactions Plot**. This procedure requires stacked data. For example, to construct the Figure 11.17 cell means plot for mobile electronic sales based on mobile payments permitted and in-store location shown on page 426, open to the **Mobile**

Electronics2 worksheet. Select **Stat → ANOVA → Interactions Plot**. In the Interactions Plot dialog box (shown below):

1. Double-click **C3 Sales** in the variables list to add **Sales** to the **Responses** box and press **Tab**.
2. Enter **C2-C3** to the **Factors** box.
3. Clear **Display full interaction plot matrix**.
4. Click **OK**.



CHAPTER 12

CONTENTS

- 12.1 Chi-Square Test for the Difference Between Two Proportions
- 12.2 Chi-Square Test for Differences Among More Than Two Proportions
- 12.3 Chi-Square Test of Independence
- 12.4 Wilcoxon Rank Sum Test: A Nonparametric Method for Two Independent Populations
- 12.5 Kruskal-Wallis Rank Test: A Nonparametric Method for the One-Way ANOVA
- 12.6 McNemar Test for the Difference Between Two Proportions (Related Samples) (*online*)
- 12.7 Chi-Square Test for the Variance or Standard Deviation (*online*)
- 12.8 Wilcoxon Signed Ranks Test: A Nonparametric Test for Two Related Populations (*online*)
- 12.9 Friedman Rank Test: A Nonparametric Test for The Randomized Block Design (*online*)

USING STATISTICS: Avoiding Guesswork About Resort Guests, Revisited

CHAPTER 12 EXCEL GUIDE

CHAPTER 12 MINITAB GUIDE

OBJECTIVES

- To learn when to use the chi-square test for contingency tables
- To learn to use the Marascuilo procedure for determining pairwise differences when evaluating more than two proportions
- To learn to use nonparametric tests

Chi-Square and Nonparametric Tests

USING STATISTICS

Avoiding Guesswork About Resort Guests

You are the manager of T.C. Resort Properties, a collection of five upscale hotels located on two tropical islands. Guests who are satisfied with the quality of services during their stay are more likely to return on a future vacation and to recommend the hotel to friends and relatives. You have defined the business objective as improving the percentage of guests who choose to return to the hotels later. To assess the quality of services being provided by your hotels, your staff encourages guests to complete a satisfaction survey when they check out or via email after they check out.

You need to analyze the data from these surveys to determine the overall satisfaction with the services provided, the likelihood that the guests will return to the hotel, and the reasons some guests indicate that they will not return. For example, on one island, T.C. Resort Properties operates the Beachcomber and Windsurfer hotels. Is the perceived quality at the Beachcomber Hotel the same as at the Windsurfer Hotel? If there is a difference, how can you use this information to improve the overall quality of service at T.C. Resort Properties? Furthermore, if guests indicate that they are not planning to return, what are the most common reasons cited for this decision? Are the reasons cited unique to a certain hotel or common to all hotels operated by T.C. Resort Properties?



Matueros1812/Shutterstock

In the preceding three chapters, you used hypothesis-testing procedures to analyze both numerical and categorical data. Chapter 9 presented some one-sample tests, Chapter 10 developed several two-sample tests, and Chapter 11 discussed analysis of variance (ANOVA). This chapter extends hypothesis testing to analyze differences between population *proportions* based on two or more samples and to test the hypothesis of independence in the joint responses to two categorical variables. The chapter concludes with nonparametric tests as alternatives to several Chapter 10 and 11 hypothesis tests.

12.1 Chi-Square Test for the Difference Between Two Proportions

In Section 10.3, you studied the *Z* test for the difference between two proportions. In this section, the differences between two proportions are examined from a different perspective. The hypothesis-testing procedure uses a test statistic, whose sampling distribution is approximated by a chi-square (χ^2) distribution. The results of this χ^2 test are equivalent to those of the *Z* test described in Section 10.3.

If you are interested in comparing the counts of categorical responses between two independent groups, you can develop a **two-way contingency table** to display the frequency of occurrence of items of interest and items not of interest for each group. (Contingency tables were first discussed in Section 2.1, and in Chapter 4, contingency tables were used to define and study probability.)

To illustrate a contingency table, return to the Using Statistics scenario concerning T.C. Resort Properties. On one of the islands, T.C. Resort Properties has two hotels (the Beachcomber and the Windsurfer). You collect data from customer satisfaction surveys and focus on the responses to the single question “Are you likely to choose this hotel again?” You organize the results of the survey and determine that 163 of 227 guests at the Beachcomber responded yes to “Are you likely to choose this hotel again?” and 154 of 262 guests at the Windsurfer responded yes to “Are you likely to choose this hotel again?” You want to analyze the results to determine whether, at the 0.05 level of significance, there is evidence of a significant difference in guest satisfaction (as measured by likelihood to return to the hotel) between the two hotels.

The contingency table displayed in Table 12.1, which has two rows and two columns, is called a **2×2 contingency table**. The cells in the table indicate the frequency for each row-and-column combination.

TABLE 12.1

Layout of a 2×2 Contingency Table

ROW VARIABLE	COLUMN VARIABLE		
	Group 1	Group 2	Totals
Items of interest	X_1	X_2	X
Items not of interest	$n_1 - X_1$	$n_2 - X_2$	$n - X$
Totals	n_1	n_2	n

where

X_1 = number of items of interest in group 1

X_2 = number of items of interest in group 2

$n_1 - X_1$ = number of items that are not of interest in group 1

$n_2 - X_2$ = number of items that are not of interest in group 2

$X = X_1 + X_2$, the total number of items of interest

$n - X = (n_1 - X_1) + (n_2 - X_2)$, the total number of items that are not of interest

$$\begin{aligned}n_1 &= \text{sample size in group 1} \\n_2 &= \text{sample size in group 2} \\n &= n_1 + n_2 = \text{total sample size}\end{aligned}$$

Table 12.2 is the contingency table for the hotel guest satisfaction study. The contingency table has two rows, indicating whether the guests would return to the hotel or would not return to the hotel, and two columns, one for each hotel. The cells in the table indicate the frequency of each row-and-column combination. The row totals indicate the number of guests who would return to the hotel and the number of guests who would not return to the hotel. The column totals are the sample sizes for each hotel location.

TABLE 12.2

2×2 Contingency Table for the Hotel Guest Satisfaction Survey

CHOOSE HOTEL AGAIN?	HOTEL		Total
	Beachcomber	Windsurfer	
Yes	163	154	317
No	64	108	172
Total	227	262	489

Student Tip

Do not confuse this use of the Greek letter pi, π , to represent the population proportion with the mathematical constant that is the ratio of the circumference to a diameter of a circle—approximately 3.14159—which is also known by the same Greek letter.

To test whether the population proportion of guests who would return to the Beachcomber, π_1 , is equal to the population proportion of guests who would return to the Windsurfer, π_2 , you can use the **chi-square (χ^2) test for the difference between two proportions**. To test the null hypothesis that there is no difference between the two population proportions:

$$H_0: \pi_1 = \pi_2$$

against the alternative that the two population proportions are not the same:

$$H_1: \pi_1 \neq \pi_2$$

you use the χ^2_{STAT} test statistic, shown in Equation (12.1) whose sampling distribution follows the chi-square distribution.

χ^2 TEST FOR THE DIFFERENCE BETWEEN TWO PROPORTIONS

The χ^2_{STAT} test statistic is equal to the squared difference between the observed and expected frequencies, divided by the expected frequency in each cell of the table, summed over all cells of the table.

$$\chi^2_{STAT} = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e} \quad (12.1)$$

where

f_o = **observed frequency** in a particular cell of a contingency table

f_e = **expected frequency** in a particular cell if the null hypothesis is true

The χ^2_{STAT} test statistic approximately follows a chi-square distribution with 1 degree of freedom.¹

¹In general, the degrees of freedom in a contingency table are equal to (number of rows – 1) multiplied by (number of columns – 1).

To compute the expected frequency, f_e , in any cell, you need to understand that if the null hypothesis is true, the proportion of items of interest in the two populations will be equal. In such situations, the sample proportions you compute from each of the two groups would differ from each other only by chance. Each would provide an estimate of the common population

Student Tip

Remember, the sample proportion, p , must be between 0 and 1.

parameter, π . A statistic that combines these two separate estimates together into one overall estimate of the population parameter provides more information than either of the two separate estimates could provide by itself. This statistic, given by the symbol \bar{p} , represents the estimated overall proportion of items of interest for the two groups combined (i.e., the total number of items of interest divided by the total sample size). The complement of \bar{p} , $1 - \bar{p}$, represents the estimated overall proportion of items that are not of interest in the two groups. Using the notation presented in Table 12.1 on page 448, Equation (12.2) defines \bar{p} .

COMPUTING THE ESTIMATED OVERALL PROPORTION FOR TWO GROUPS

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{X}{n} \quad (12.2)$$

Student Tip

Remember that the rejection region for this test is only in the upper tail of the chi-square distribution.

To compute the expected frequency, f_e , for cells that involve items of interest (i.e., the cells in the first row in the contingency table), you multiply the sample size (or column total) for a group by \bar{p} . To compute the expected frequency, f_e , for cells that involve items that are not of interest (i.e., the cells in the second row in the contingency table), you multiply the sample size (or column total) for a group by $1 - \bar{p}$.

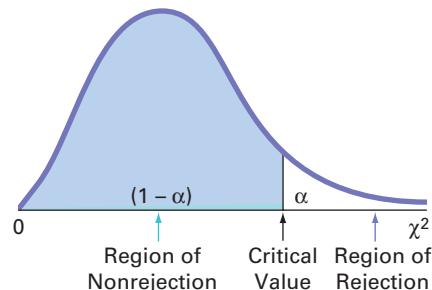
The sampling distribution of the χ^2_{STAT} test statistic shown in Equation (12.1) on page 449 approximately follows a **chi-square (χ^2) distribution** (see Table E.4) with 1 degree of freedom. Using a level of significance α , you reject the null hypothesis if the computed χ^2_{STAT} test statistic is greater than χ^2_α , the upper-tail critical value from the χ^2 distribution with 1 degree of freedom. Thus, the decision rule is

$$\begin{aligned} \text{Reject } H_0 \text{ if } \chi^2_{STAT} &> \chi^2_\alpha; \\ \text{otherwise, do not reject } H_0. \end{aligned}$$

Figure 12.1 illustrates the decision rule.

FIGURE 12.1

Regions of rejection and nonrejection when using the chi-square test for the difference between two proportions, with level of significance α



If the null hypothesis is true, the computed χ^2_{STAT} test statistic should be close to zero because the squared difference between what is actually observed in each cell, f_o , and what is theoretically expected, f_e , should be very small. If H_0 is false, then there are differences in the population proportions, and the computed χ^2_{STAT} test statistic is expected to be large. However, what is a large difference in a cell is relative. Because you are dividing by the expected frequencies, the same actual difference between f_o and f_e from a cell with a small number of expected frequencies contributes more to the χ^2_{STAT} test statistic than a cell with a large number of expected frequencies.

To illustrate the use of the chi-square test for the difference between two proportions, return to the Using Statistics scenario concerning T.C. Resort Properties on page 447 and the corresponding contingency table displayed in Table 12.2 on page 449. The null hypothesis

$(H_0: \pi_1 = \pi_2)$ states that there is no difference between the proportion of guests who are likely to choose either of these hotels again. To begin,

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{163 + 154}{227 + 262} = \frac{317}{489} = 0.6483$$

\bar{p} is the estimate of the common parameter π , the population proportion of guests who are likely to choose either of these hotels again if the null hypothesis is true. The estimated proportion of guests who are *not* likely to choose these hotels again is the complement of \bar{p} , $1 - 0.6483 = 0.3517$. Multiplying these two proportions by the sample size for the Beachcomber Hotel gives the number of guests expected to choose the Beachcomber again and the number not expected to choose this hotel again. In a similar manner, multiplying the two proportions by the Windsurfer Hotel's sample size yields the corresponding expected frequencies for that group.

EXAMPLE 12.1

Computing the Expected Frequencies

Compute the expected frequencies for each of the four cells of Table 12.2 on page 449.

SOLUTION

Yes—Beachcomber: $\bar{p} = 0.6483$ and $n_1 = 227$, so $f_e = 147.16$

Yes—Windsurfer: $\bar{p} = 0.6483$ and $n_2 = 262$, so $f_e = 169.84$

No—Beachcomber: $1 - \bar{p} = 0.3517$ and $n_1 = 227$, so $f_e = 79.84$

No—Windsurfer: $1 - \bar{p} = 0.3517$ and $n_2 = 262$, so $f_e = 92.16$

Table 12.3 presents these expected frequencies next to the corresponding observed frequencies.

TABLE 12.3

Comparing the Observed (f_o) and Expected (f_e) Frequencies

CHOOSE HOTEL AGAIN?	HOTEL				Total	
	Beachcomber		Windsurfer			
	Observed	Expected	Observed	Expected		
Yes	163	147.16	154	169.84	317	
No	64	79.84	108	92.16	172	
Total	227	227.00	262	262.00	489	

To test the null hypothesis that the population proportions are equal:

$$H_0: \pi_1 = \pi_2$$

against the alternative that the population proportions are not equal:

$$H_1: \pi_1 \neq \pi_2$$

you use the observed and expected frequencies from Table 12.3 to compute the χ^2_{STAT} test statistic given by Equation (12.1) on page 449. Table 12.4 presents these calculations.

TABLE 12.4

Computing the χ^2_{STAT} Test Statistic for the Hotel Guest Satisfaction Survey

f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
163	147.16	15.84	250.91	1.71
154	169.84	-15.84	250.91	1.48
64	79.84	-15.84	250.91	3.14
108	92.16	15.84	250.91	2.72
				9.05

The chi-square (χ^2) distribution is a right-skewed distribution whose shape depends solely on the number of degrees of freedom. You find the critical value for the χ^2 test from Table E.4, a portion of which is presented in Table 12.5.

TABLE 12.5

Finding the Critical Value from the Chi-Square Distribution with 1 Degree of Freedom, Using the 0.05 Level of Significance

Degrees of Freedom	Cumulative Probabilities					
	.005	.0195	.975	.99
	Upper-Tail Area					
1	.995	.9905	.025	.01
2		0.010	0.020	...	5.991	7.378
3		0.072	0.115	...	7.815	9.348
4		0.207	0.297	...	9.488	11.143
5		0.412	0.554	...	11.071	12.833
						15.086
						16.750

The values in Table 12.5 refer to selected upper-tail areas of the χ^2 distribution. A 2×2 contingency table has 1 degree of freedom because there are two rows and two columns. [The degrees of freedom are equal to the (number of rows - 1)(number of columns - 1).] Using $\alpha = 0.05$, with 1 degree of freedom, the critical value of χ^2 from Table 12.5 is 3.841. You reject H_0 if the computed χ^2_{STAT} test statistic is greater than 3.841 (see Figure 12.2). Because $\chi^2_{STAT} = 9.05 > 3.841$, you reject H_0 . You conclude that the proportion of guests who would return to the Beachcomber is different from the proportion of guests who would return to the Windsurfer.

FIGURE 12.2

Regions of rejection and nonrejection when finding the χ^2 critical value with 1 degree of freedom, at the 0.05 level of significance

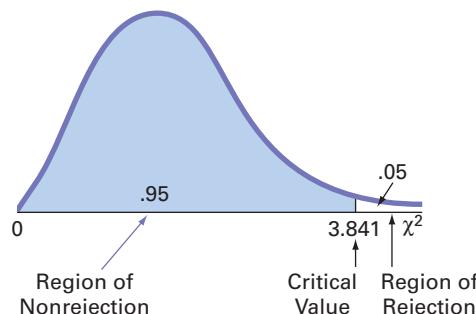


Figure 12.3 shows the Excel and Minitab results for the Table 12.2 guest satisfaction contingency table on page 449.

FIGURE 12.3

Excel and Minitab results of the chi-square test for the two-hotel guest satisfaction survey

Chi-Square Test						Chi-Square Test: Beachcomber, Windsurfer			
						Expected counts are printed below observed counts Chi-Square contributions are printed below expected counts			
Observed Frequencies						Beachcomber	Windsurfer	Total	
Choose Again?	Beachcomber	Windsurfer	Total			1	163	154	317
Yes	163	154	317	15.8446	-15.8446		147.16	169.84	
No	64	108	172	-15.8446	15.8446		1.706	1.478	
Total	227	262	489			2	64	108	172
							79.84	92.16	
							3.144	2.724	
						Total	227	262	489
Expected Frequencies						Chi-Sq = 9.053, DF = 1, P-Value = 0.003			
Choose Again?	Beachcomber	Windsurfer	Total						
Yes	147.1554	169.8446	317	1.7060	1.4781				
No	79.8446	92.1554	172	3.1442	2.7242				
Total	227	262	489						
Data									
Level of Significance	0.05								
Number of Rows	2								
Number of Columns	2								
Degrees of Freedom	1								
						= (B19 - 1) * (B20 - 1)			
Results									
Critical Value	3.8415					=CHISQ.INV.RT(B18, B21)			
Chi-Square Test Statistic	9.0526					=SUM(F13:G14)			
p-Value	0.0026					=CHISQ.DIST.RT(B25, B21)			
Reject the null hypothesis						=IF(B26 < B18, "Reject the null hypothesis", "Do not reject the null hypothesis")			
Expected frequency assumption is met.						=IF(OR(B13 < 5, C13 < 5, B14 < 5, C14 < 5), " is violated.", " is met.")			

These results include the expected frequencies, χ_{STAT}^2 , degrees of freedom, and p -value. The computed χ_{STAT}^2 test statistic is 9.0526, which is greater than the critical value of 3.8415 (or the p -value = 0.0026 < 0.05), so you reject the null hypothesis that there is no difference in guest satisfaction between the two hotels. The p -value, equal to 0.0026, is the probability of observing sample proportions as different as or more different from the actual difference between the Beachcomber and Windsurfer ($0.718 - 0.588 = 0.13$) observed in the sample data, if the population proportions for the Beachcomber and Windsurfer hotels are equal. Thus, there is strong evidence to conclude that the two hotels are significantly different with respect to guest satisfaction, as measured by whether a guest is likely to return to the hotel again. From Table 12.3 on page 451 you can see that a greater proportion of guests are likely to return to the Beachcomber than to the Windsurfer.

For the χ^2 test to give accurate results for a 2×2 table, you must assume that each expected frequency is at least 5. If this assumption is not satisfied, you can use alternative procedures, such as Fisher's exact test (see references 1, 2, and 4).

In the hotel guest satisfaction survey, both the Z test based on the standardized normal distribution (see Section 10.3) and the χ^2 test based on the chi-square distribution lead to the same conclusion. You can explain this result by the interrelationship between the standardized normal distribution and a chi-square distribution with 1 degree of freedom. For such situations, the χ_{STAT}^2 test statistic is the square of the Z_{STAT} test statistic.

For example, in the guest satisfaction study, the computed Z_{STAT} test statistic is +3.0088, and the computed χ_{STAT}^2 test statistic is 9.0526. Except for rounding differences, this 9.0526 value is the square of +3.0088 [i.e., $(+3.0088)^2 \approx 9.0526$]. Also, if you compare the critical values of the test statistics from the two distributions, at the 0.05 level of significance, the χ^2 value of 3.841 with 1 degree of freedom is the square of the Z value of ± 1.96 . Furthermore, the p -values for both tests are equal. Therefore, when testing the null hypothesis of equality of proportions:

$$H_0: \pi_1 = \pi_2$$

against the alternative that the population proportions are not equal:

$$H_1: \pi_1 \neq \pi_2$$

the Z test and the χ^2 test are equivalent. If you are interested in determining whether there is evidence of a *directional* difference, such as $\pi_1 > \pi_2$, you must use the Z test, with the entire rejection region located in one tail of the standardized normal distribution.

In Section 12.2, the χ^2 test is extended to make comparisons and evaluate differences between the proportions among more than two groups. However, you cannot use the Z test if there are more than two groups.

Problems for Section 12.1

LEARNING THE BASICS

12.1 Determine the critical value of χ^2 with 1 degree of freedom in each of the following circumstances:

- a. $\alpha = 0.01$
- b. $\alpha = 0.005$
- c. $\alpha = 0.10$

12.2 Determine the critical value of χ^2 with 1 degree of freedom in each of the following circumstances:

- a. $\alpha = 0.05$
- b. $\alpha = 0.025$
- c. $\alpha = 0.01$

12.3 Use the following contingency table:

	A	B	Total
1	20	30	50
2	30	45	75
Total	50	75	125

- a. Compute the expected frequency for each cell.
- b. Compare the observed and expected frequencies for each cell.
- c. Compute χ_{STAT}^2 . Is it significant at $\alpha = 0.05$?

12.4 Use the following contingency table:

	A	B	Total
1	20	30	50
2	30	20	50
Total	50	50	100

- a. Compute the expected frequency for each cell.
 b. Compute χ^2_{STAT} . Is it significant at $\alpha = 0.05$?

APPLYING THE CONCEPTS

12.5 A survey of 1,085 adults asked, “Do you enjoy shopping for clothing for yourself?” The results indicated that 51% of the females enjoyed shopping for clothing for themselves as compared to 44% of the males. (Data extracted from “Split Decision on Clothes Shopping,” *USA Today*, January 28, 2011, p. 1B.) The sample sizes of males and females were not provided. Suppose that the results were as shown in the following table:

ENJOY SHOPPING FOR CLOTHING	GENDER		Total
	Male	Female	
Yes	238	276	514
No	304	267	571
Total	542	543	1,085

- a. Is there evidence of a significant difference between the proportion of males and females who enjoy shopping for clothing for themselves at the 0.01 level of significance?
 b. Determine the p -value in (a) and interpret its meaning.
 c. What are your answers to (a) and (b) if 218 males enjoyed shopping for clothing and 324 did not?
 d. Compare the results of (a) through (c) to those of Problem 10.29 (a), (b), and (d) on page 372.

12.6 Do social recommendations increase ad effectiveness? A study of online video viewers compared viewers who arrived at an advertising video for a particular brand by following a social media recommendation link to viewers who arrived at the same video by web browsing. Data were collected on whether the viewer could correctly recall the brand being advertised after seeing the video. The results were:

ARRIVAL METHOD	CORRECTLY RECALLED THE BRAND	
	Yes	No
Recommendation	407	150
Browsing	193	91

Source: Data extracted from “Social Ad Effectiveness: An Unruly White Paper,” www.unrulymedia.com, January 2012, p. 3.

- a. Set up the null and alternative hypotheses to determine whether there is a difference in brand recall between viewers who arrived by following a social media recommendation and those who arrived by web browsing.

- b. Conduct the hypothesis test defined in (a), using the 0.05 level of significance.
 c. Compare the results of (a) and (b) to those of Problem 10.30 (a) and (b) on page 372.

12.7 Do males or females feel more tense or stressed out at work? A survey of employed adults conducted online by Harris Interactive on behalf of the American Psychological Association revealed the following:

FELT TENSE OR STRESSED OUT AT WORK		
GENDER	Yes	No
Male	244	495
Female	282	480

Source: Data extracted from “The 2013 Work and Well-Being Survey,” American Psychological Association and Harris Interactive, March 2013, p. 5, bit.ly/11JGePf.

- a. At the 0.05 level of significance, is there evidence of a difference between males and females in the proportion who feel stressed out at work?
 b. Determine the p -value in (a) and interpret its meaning.

- SELF TEST** **12.8** Consumer research firm Scarborough analyzed the 10% of American adults who are either “Superbanked” or “Unbanked.” Superbanked consumers are defined as U.S. adults who live in a household that has multiple asset accounts at financial institutions, as well as some additional investments; Unbanked consumers are U.S. adults who live in a household that does not use a bank or credit union. By finding the 5% of Americans who are Superbanked, Scarborough identifies financially savvy consumers who might be open to diversifying their financial portfolios; by identifying the Unbanked, Scarborough provides insight into the ultimate prospective client for banks and financial institutions. As part of its analysis, Scarborough reported that 93% of Superbanked consumers use credit cards as compared to 23% of Unbanked consumers. (Data extracted from bit.ly/Sy19kN.) Suppose that these results were based on 1,000 Superbanked consumers and 1,000 Unbanked consumers.
- a. At the 0.01 level of significance, is there evidence of a significant difference between the Superbanked and the Unbanked with respect to the proportion that use credit cards?
 b. Determine the p -value in (a) and interpret its meaning.
 c. Compare the results of (a) and (b) to those of Problem 10.32 on page 373.

12.9 A/B testing is a method used by businesses to test different designs and formats of a web page to determine if a new web page is more effective than a current web page. Web designers tested a new call to action button on its web page. Every visitor to the web page was randomly shown either the original call-to-action button (the control) or the new variation. The metric used to measure success was the download rate: the number of people who downloaded the file divided by the number of people who saw that particular call-to-action button. Results of the experiment yielded the following:

VARIATIONS	DOWNLOADS	VISITORS
Original call-to-action button	351	3,642
New call-to-action button	485	3,556

- At the 0.05 level of significance, is there evidence of a difference in the download rate between the original call to action button and the new call to action button?
- Find the p -value in (a) and interpret its value.
- Compare the results of (a) and (b) to those of Problem 10.31 on page 372.

12.10 Does gamification motivate customer research management (CRM) utilization? Gamification is the use of game mechanics to motivate, modify, or reward distinct behaviors. In the context of sales effectiveness, it is deployed to encourage both sales accomplishments and nonsales activities. A survey of end-user sales organizations indicates that 31 of 37 gamification-user

organizations provide mobile access to CRM, whereas 138 of 275 non-gamification-user organizations provide mobile access to CRM. (Data extracted from bit.ly/12Omho7.)

- Construct a 2×2 contingency table.
- At the 0.05 level of significance, is there evidence of a difference between gamification-user sales organizations and non-gamification-user sales organizations in the proportion that provide mobile access to CRM?
- Find the p -value in (a) and interpret its meaning.
- Compare the results of (a) and (b) to those of Problem 10.34 on page 373.

12.2 Chi-Square Test for Differences Among More Than Two Proportions

In this section, the χ^2 test is extended to compare more than two independent populations. The letter c is used to represent the number of independent populations under consideration. Thus, the contingency table now has two rows and c columns. To test the null hypothesis that there are no differences among the c population proportions:

$$H_0: \pi_1 = \pi_2 = \cdots = \pi_c$$

against the alternative that not all the c population proportions are equal:

$$H_1: \text{Not all } \pi_j \text{ are equal (where } j = 1, 2, \dots, c\text{)}$$

you use Equation (12.1) on page 449:

$$\chi_{STAT}^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

where

f_o = observed frequency in a particular cell of a $2 \times c$ contingency table

f_e = expected frequency in a particular cell if the null hypothesis is true

If the null hypothesis is true and the proportions are equal across all c populations, the c sample proportions should differ only by chance. In such a situation, a statistic that combines these c separate estimates into one overall estimate of the population proportion, π , provides more information than any one of the c separate estimates alone. To expand on Equation (12.2) on page 450, the statistic \bar{p} in Equation (12.3) represents the estimated overall proportion for all c groups combined.

COMPUTING THE ESTIMATED OVERALL PROPORTION FOR c GROUPS

$$\bar{p} = \frac{X_1 + X_2 + \cdots + X_c}{n_1 + n_2 + \cdots + n_c} = \frac{X}{n} \quad (12.3)$$

To compute the expected frequency, f_e , for each cell in the first row in the contingency table, multiply each sample size (or column total) by \bar{p} . To compute the expected frequency, f_e , for each cell in the second row in the contingency table, multiply each sample size (or column total) by $(1 - \bar{p})$. The sampling distribution of the test statistic shown in Equation (12.1) on page 449 approximately follows a chi-square distribution, with degrees of freedom equal to the number of rows in the contingency table minus 1, multiplied by the number of columns in the table minus 1. For a $2 \times c$ contingency table, there are $c - 1$ degrees of freedom:

$$\text{Degrees of freedom} = (2 - 1)(c - 1) = c - 1$$

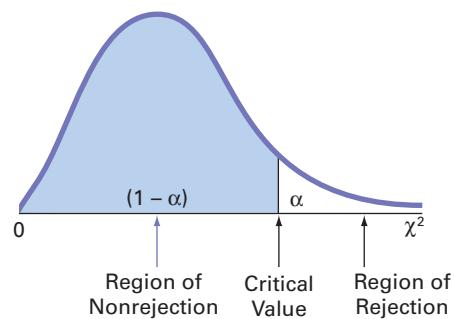
Using the level of significance α , you reject the null hypothesis if the computed χ^2_{STAT} test statistic is greater than χ^2_α , the upper-tail critical value from a chi-square distribution with $c - 1$ degrees of freedom. Therefore, the decision rule is

$$\text{Reject } H_0 \text{ if } \chi^2_{STAT} > \chi^2_\alpha; \\ \text{otherwise, do not reject } H_0.$$

Figure 12.4 illustrates this decision rule.

FIGURE 12.4

Regions of rejection and nonrejection when testing for differences among c proportions using the χ^2 test



To illustrate the χ^2 test for equality of proportions when there are more than two groups, return to the Using Statistics scenario on page 447 concerning T.C. Resort Properties. Once again, you define the business objective as improving the quality of service, but this time, you are comparing three hotels located on a different island. Data are collected from customer satisfaction surveys at these three hotels. You organize the responses into the contingency table shown in Table 12.6.

TABLE 12.6

2 × 3 Contingency Table for Guest Satisfaction Survey

CHOOSE HOTEL AGAIN?	HOTEL			Total
	Golden Palm	Palm Royale	Palm Princess	
Yes	128	199	186	513
No	88	33	66	187
Total	216	232	252	700

Because the null hypothesis states that there are no differences among the three hotels in the proportion of guests who would likely return again, you use Equation (12.3) to calculate an estimate of π , the population proportion of guests who would likely return again:

$$\begin{aligned}\bar{p} &= \frac{X_1 + X_2 + \cdots + X_c}{n_1 + n_2 + \cdots + n_c} = \frac{X}{n} \\ &= \frac{(128 + 199 + 186)}{(216 + 232 + 252)} = \frac{513}{700} \\ &= 0.733\end{aligned}$$

The estimated overall proportion of guests who would *not* be likely to return again is the complement, $(1 - \bar{p})$, or 0.267. Multiplying these two proportions by the sample size for each hotel yields the expected number of guests who would and would not likely return.

EXAMPLE 12.2

Computing the
Expected
Frequencies

Compute the expected frequencies for each of the six cells in Table 12.6.

SOLUTION

Yes—Golden Palm: $\bar{p} = 0.733$ and $n_1 = 216$, so $f_e = 158.30$

Yes—Palm Royale: $\bar{p} = 0.733$ and $n_2 = 232$, so $f_e = 170.02$

Yes—Palm Princess: $\bar{p} = 0.733$ and $n_3 = 252$, so $f_e = 184.68$

No—Golden Palm: $1 - \bar{p} = 0.267$ and $n_1 = 216$, so $f_e = 57.70$

No—Palm Royale: $1 - \bar{p} = 0.267$ and $n_2 = 232$, so $f_e = 61.98$

No—Palm Princess: $1 - \bar{p} = 0.267$ and $n_3 = 252$, so $f_e = 67.32$

Table 12.7 presents these expected frequencies.

TABLE 12.7

Contingency Table of
Expected Frequencies
from a Guest
Satisfaction Survey
of Three Hotels

CHOOSE HOTEL AGAIN?	HOTEL			Total
	Golden Palm	Palm Royale	Palm Princess	
Yes	158.30	170.02	184.68	513
No	57.70	61.98	67.32	187
Total	216.00	232.00	252.00	700

To test the null hypothesis that the proportions are equal:

$$H_0: \pi_1 = \pi_2 = \pi_3$$

against the alternative that not all three proportions are equal:

$$H_1: \text{Not all } \pi_j \text{ are equal (where } j = 1, 2, 3\text{)}$$

you use the observed frequencies from Table 12.6 and the expected frequencies from Table 12.7 to compute the χ^2_{STAT} test statistic [given by Equation (12.1) on page 449]. Table 12.8 presents the calculations.

TABLE 12.8

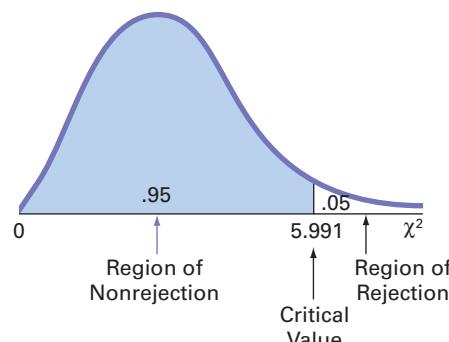
Computing the χ^2_{STAT}
Test Statistic for the
Three-Hotel Guest
Satisfaction Survey

f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
128	158.30	-30.30	918.09	5.80
199	170.02	28.98	839.84	4.94
186	184.68	1.32	1.74	0.01
88	57.70	30.30	918.09	15.91
33	61.98	-28.98	839.84	13.55
66	67.32	-1.32	1.74	0.02
				40.23

You use Table E.4 to find the critical value of the χ^2 test statistic. In the guest satisfaction survey, because there are three hotels, there are $(2 - 1)(3 - 1) = 2$ degrees of freedom. Using $\alpha = 0.05$, the χ^2 critical value with 2 degrees of freedom is 5.991 (see Figure 12.5).

FIGURE 12.5

Regions of rejection and
nonrejection when testing
for differences in three
proportions at the
0.05 level of significance,
with 2 degrees of
freedom



Because the computed χ^2_{STAT} test statistic is 40.23, which is greater than this critical value, you reject the null hypothesis. Figure 12.6 shows the Excel and Minitab results for this problem. These results also report the p -value. Because the p -value is 0.0000, less than $\alpha = 0.05$, you reject the null hypothesis. Further, this p -value indicates that there is virtually no chance that there will be differences this large or larger among the three sample proportions, if the population proportions for the three hotels are equal. Thus, there is sufficient evidence to conclude that the hotel properties are different with respect to the proportion of guests who are likely to return.

FIGURE 12.6

Excel and Minitab chi-square test results for the three-hotel guest satisfaction survey

Chi-Square Test								
Observed Frequencies								
Choose Again?	Hotel			Calculations				
	Golden Palm	Palm Royale	Palm Princess	fo - fe				
Yes	128	199	186	513	-30.2971	28.9771	1.32	
No	88	33	66	187	30.2971	-28.9771	-1.32	
Total	216	232	252	700				
Expected Frequencies								
Choose Again?	Hotel			(fo - fe)^2/fe				
	Golden Palm	Palm Royale	Palm Princess	Total	5.7987	4.9386	0.0094	
Yes	158.2971	170.0229	184.68	513	5.7987	4.9386	0.0094	
No	57.7029	61.9771	67.32	187	15.9077	13.5481	0.0259	
Total	216	232	252	700				
Data								
Level of Significance	0.05							
Number of Rows	2							
Number of Columns	3							
Degrees of Freedom	2							
Results								
Critical Value	5.9915							
Chi-Square Test Statistic	=B19 - 1) * (B20 - 1)							
p-Value	=CHISQ.INV.RT(B18, B21)							
Reject the null hypothesis	=SUM(G13:I14)							
Expected frequency assumption is met.	=CHISQ.DIST.RT(B25, B21)							
	=IF(B26 < B18, "Reject the null hypothesis", "Do not reject the null hypothesis")							

Chi-Square Test: Golden Palm, Palm Royale, Palm Princess				
Expected counts are printed below observed counts				
Chi-Square contributions are printed below expected counts				
Golden	Palm	Palm	Total	
1	128	199	186	513
	158.30	170.02	184.68	
	5.799	4.939	0.009	
2	88	33	66	187
	57.70	61.98	67.32	
	15.908	13.548	0.026	
Total	216	232	252	700
Chi-Sq = 40.228, DF = 2, P-Value = 0.000				

For the χ^2 test to give accurate results when dealing with $2 \times c$ contingency tables, all expected frequencies must be large. The definition of “large” has led to research among statisticians. Some statisticians (see reference 5) have found that the test gives accurate results as long as all expected frequencies are at least 0.5. Other statisticians, more conservative in their approach, believe that no more than 20% of the cells should contain expected frequencies less than 5, and no cells should have expected frequencies less than 1 (see reference 3). As a reasonable compromise between these points of view, to ensure the validity of the test, you should make sure that each expected frequency is at least 1. To do this, you may need to collapse two or more low-expected-frequency categories into one category in the contingency table before performing the test. If combining categories is undesirable, you can use one of the available alternative procedures (see references 1, 2, and 7).

The Marascuilo Procedure

Rejecting the null hypothesis in a χ^2 test of equality of proportions in a $2 \times c$ table allows you to only reach the conclusion that not all c population proportions are equal. To determine which proportions differ, you use a multiple-comparisons procedure such as the Marascuilo procedure.

The **Marascuilo procedure** enables you to make comparisons between all pairs of groups. First, you compute the sample proportions. Then, you use Equation (12.4) to compute the critical ranges for the Marascuilo procedure. You compute a different critical range for each pairwise comparison of sample proportions.

CRITICAL RANGE FOR THE MARASCUILO PROCEDURE

$$\text{Critical range} = \sqrt{\chi_{\alpha}^2} \sqrt{\frac{p_j(1 - p_j)}{n_j} + \frac{p_{j'}(1 - p_{j'})}{n_{j'}}} \quad (12.4)$$

where

p_j = proportion of items of interest in group j

$p_{j'}$ = proportion of items of interest in group j'

n_j = sample size in group j

$n_{j'}$ = sample size in group j'

Student Tip

You have an α level of risk in the entire set of comparisons not just a single comparison.

Then, you compare each of the $c(c - 1)/2$ pairs of sample proportions against its corresponding critical range. You declare a specific pair significantly different if the absolute difference in the sample proportions, $|p_j - p_{j'}|$, is greater than its critical range.

To apply the Marascuilo procedure, return to the three-hotel guest satisfaction survey. Using the χ^2 test, you concluded that there was evidence of a significant difference among the population proportions. From Table 12.6 on page 456, the three sample proportions are

$$p_1 = \frac{X_1}{n_1} = \frac{128}{216} = 0.5926$$

$$p_2 = \frac{X_2}{n_2} = \frac{199}{232} = 0.8578$$

$$p_3 = \frac{X_3}{n_3} = \frac{186}{252} = 0.7381$$

Next, you compute the absolute differences in sample proportions and their corresponding critical ranges. Because there are three hotels, there are $(3)(3 - 1)/2 = 3$ pairwise comparisons. Using Table E.4 and an overall level of significance of 0.05, the upper-tail critical value for a chi-square distribution having $(c - 1) = 2$ degrees of freedom is 5.991. Thus,

$$\sqrt{\chi_{\alpha}^2} = \sqrt{5.991} = 2.4477$$

The following displays the absolute differences and the critical ranges for each comparison.

Absolute Difference in Proportions	Critical Range
$ p_j - p_{j'} $	$2.4477 \sqrt{\frac{p_j(1 - p_j)}{n_j} + \frac{p_{j'}(1 - p_{j'})}{n_{j'}}}$
$ p_1 - p_2 = 0.5926 - 0.8578 = 0.2652$	$2.4477 \sqrt{\frac{(0.5926)(0.4074)}{216} + \frac{(0.8578)(0.1422)}{232}} = 0.0992$
$ p_1 - p_3 = 0.5926 - 0.7381 = 0.1455$	$2.4477 \sqrt{\frac{(0.5926)(0.4074)}{216} + \frac{(0.7381)(0.2619)}{252}} = 0.1063$
$ p_2 - p_3 = 0.8578 - 0.7381 = 0.1197$	$2.4477 \sqrt{\frac{(0.8578)(0.1422)}{232} + \frac{(0.7381)(0.2619)}{252}} = 0.0880$

Figure 12.7 shows Excel results for this example.

FIGURE 12.7

Excel Marascuilo procedure results for the three-hotel guest satisfaction survey

A	B	C	D
Marascuilo Procedure for Guest Satisfaction Analysis			
3	Level of Significance	0.05	=ChiSquare2x3!B18
4	Square Root of Critical Value	2.4477	=SQRT(ChiSquare2x3!B24)
Group Sample Proportions			
7	1: Golden Palm	0.5926	=ChiSquare2x3!B6/ChiSquare2x3!B8
8	2: Palm Royale	0.8578	=ChiSquare2x3!C6/ChiSquare2x3!C8
9	3: Palm Princess	0.7381	=ChiSquare2x3!D6/ChiSquare2x3!D8
MARASCUILO TABLE			
12	Proportions	Absolute Differences	Critical Range
13	Group 1 - Group 2	0.2652	0.0992 Significant
14	Group 1 - Group 3	0.1455	0.1063 Significant
15			
16	Group 2 - Group 3	0.1197	0.0880 Significant

As the final step, you compare the absolute differences to the critical ranges. If the absolute difference is greater than the critical range, the proportions are significantly different. At the 0.05 level of significance, you can conclude that guest satisfaction is higher at the Palm Royale ($p_2 = 0.858$) than at either the Golden Palm ($p_1 = 0.593$) or the Palm Princess ($p_3 = 0.738$) and that guest satisfaction is also higher at the Palm Princess than at the Golden Palm. These results clearly suggest that you should investigate possible reasons for these differences. In particular, you should try to determine why satisfaction is significantly lower at the Golden Palm than at the other two hotels.

The Analysis of Proportions (ANOP)

In many situations, you need to examine differences among several proportions. The analysis of proportions (ANOP) method provides a confidence interval approach that allows you to determine which, if any, of the c groups has a proportion significantly different from the overall mean of all the group proportions combined. The **ANOP online topic** discusses this method and illustrates its use.

Problems for Section 12.2

LEARNING THE BASICS

12.11 Consider a contingency table with two rows and five columns.

- a. How many degrees of freedom are there in the contingency table?
- b. Determine the critical value for $\alpha = 0.05$.
- c. Determine the critical value for $\alpha = 0.01$.

12.12 Use the following contingency table:

	A	B	C	Total
1	10	30	50	90
2	40	45	50	135
Total	50	75	100	225

- a. Compute the expected frequency for each cell.
- b. Compute χ^2_{STAT} . Is it significant at $\alpha = 0.05$?

12.13 Use the following contingency table:

	A	B	C	Total
1	20	30	25	75
2	30	20	25	75
Total	50	50	50	150

- a. Compute the expected frequency for each cell.
- b. Compute χ^2_{STAT} . Is it significant at $\alpha = 0.05$?

APPLYING THE CONCEPTS

12.14 Do workers prefer to buy lunch rather than pack their own lunch? A survey of employed Americans found that 75% of the 18- to 24-year-olds, 77% of the 25- to 34-year-olds, 72% of the 35- to 44-year-olds, 58% of the 45- to 54-year-olds, 57% of the 54- to 64-year-olds, and 55% of the 65+ -year-olds buy lunch throughout the work week. (Data extracted from bit.ly/z99CeN.) Suppose the survey was based on 200 employed Americans in

each of six age groups: 18 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, and 65+.

- At the 0.05 level of significance, is there evidence of a difference among the age groups in the preference for buying lunch?
- Determine the p -value in (a) and interpret its meaning.
- If appropriate, use the Marascuilo procedure and $\alpha = 0.05$ to determine which age groups differ.

12.15 What are travelers' technologies of choice? Tablets account for a growing share of the multiuse devices travelers are using. An observational study of passengers on air, bus, and train travel found that 8.4% of airline passengers, 5.9% of Amtrak passengers, 4.9% of commuter train passengers, and 3.7% of curbside bus passengers were observed using a tablet at some point during their travel. (Data extracted from afterhours.e-strategy.com/passenger-tablet-use-by-transportation-mode-c.) Suppose these results were based on 500 passengers in each of the four transportation modes: airline, Amtrak train, commuter train, and curbside bus.

- At the 0.05 level of significance, is there evidence of a difference among the transportation modes with respect to use of tablets?
- Compute the p -value and interpret its meaning.
- If appropriate, use the Marascuilo procedure and $\alpha = 0.05$ to determine which transportation modes differ.

 **SELF Test** **12.16** Social media users use a variety of devices to access social networking; mobile phones are increasingly popular. However, is there a difference in the various age groups in the proportion of social media users who use their mobile phone to access social networking? A study showed the following results for the different age groups:

USE MOBILE PHONE TO ACCESS SOCIAL NETWORKING?	AGE		
	18–34	35–64	65 +
Yes	59%	36%	13%
No	41%	64%	87%

Source: Data extracted from "State of the Media: U.S. Digital Consumer Report Q3–Q4 2011," The Nielsen Company, 2012, p. 9.

Assume that 200 social media users for each age group were surveyed.

- At the 0.05 level of significance, is there evidence of a difference among the age groups with respect to use of mobile phone for accessing social networking?

- Determine the p -value in (a) and interpret its meaning.
- If appropriate, use the Marascuilo procedure and $\alpha = 0.05$ to determine which age groups differ with respect to use of a mobile phone for accessing social networking.
- Discuss the implications of (a) and (c). How can marketers use this information to improve their sales return on investment (ROI)?

12.17 Repeat (a) and (b) of Problem 12.16, assuming that only 10 social media users for each age group were surveyed. Discuss the implications of sample size on the χ^2 test for differences among more than two populations.

12.18 Who uses a cellphone while watching TV? The Pew Research Center's Internet and American Life Project measured the prevalence of multiscreen viewing experiences by asking American adults who are cellphone owners whether they had used their phone to engage in several activities while watching TV. The study reported that 171 of 316 (54%) of urban American cellphone owners sampled, 516 of 993 (52%) of suburban American cellphone owners sampled, and 251 of 557 (45%) of rural American cellphone owners used their phone to engage in several activities while watching TV. (Data extracted from "The Rise of the Connected Viewer," Pew Research Center's Internet & American Life Project, July 17, 2012.)

- Is there evidence of a significant difference among the urban, suburban, and rural American cellphone owners with respect to the proportion who use their phone to engage in several activities while watching TV? (Use $\alpha = 0.05$).
- Determine the p -value and interpret its meaning.
- If appropriate, use the Marascuilo procedure and $\alpha = 0.05$ to determine which groups differ.

12.19 The GMI Ratings' 2012 Women on Boards Survey showed incremental improvements in most measures of female board representation in the past year. The study reported that 90 of 101 (89%) of French companies sampled, 136 of 197 (69%) of Australian companies sampled, 26 of 28 (93%) of Norwegian companies sampled, 27 of 53 (51%) of Singaporean companies, and 95 of 134 (71%) of Canadian companies sampled have at least one female director on their boards. (Data extracted from bit.ly/zBAnYv.)

- Is there evidence of a significant difference among the countries with respect to the proportion of companies who have at least one female director on their boards? (Use $\alpha = 0.05$).
- Determine the p -value and interpret its meaning.
- If appropriate, use the Marascuilo procedure and $\alpha = 0.05$ to determine which countries differ.

12.3 Chi-Square Test of Independence

In Sections 12.1 and 12.2, you used the χ^2 test to evaluate potential differences among population proportions. For a contingency table that has r rows and c columns, you can generalize the χ^2 test as a *test of independence* for two categorical variables.

For a test of independence, the null and alternative hypotheses follow:

H_0 : The two categorical variables are independent (i.e., there is no relationship between them).

H_1 : The two categorical variables are dependent (i.e., there is a relationship between them).

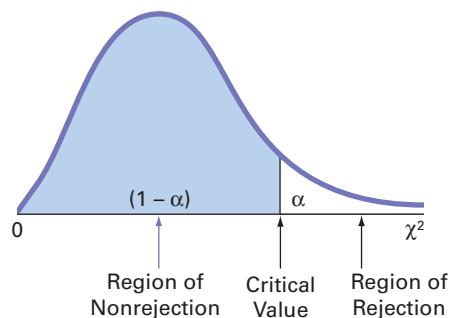
Once again, you use Equation (12.1) on page 449 to compute the test statistic:

$$\chi^2_{STAT} = \sum_{all\ cells} \frac{(f_o - f_e)^2}{f_e}$$

You reject the null hypothesis at the α level of significance if the computed value of the χ^2_{STAT} test statistic is greater than χ^2_α , the upper-tail critical value from a chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom (see Figure 12.8).

FIGURE 12.8

Regions of rejection and nonrejection when testing for independence in an $r \times c$ contingency table, using the χ^2 test



Thus, the decision rule is

Reject H_0 if $\chi^2_{STAT} > \chi^2_\alpha$;
otherwise, do not reject H_0 .

Student Tip

Remember that *independence* means no relationship, so you do not reject the null hypothesis. *Dependence* means there is a relationship, so you reject the null hypothesis.

The **chi-square (χ^2) test of independence** is similar to the χ^2 test for equality of proportions. The test statistics and the decision rules are the same, but the null and alternative hypotheses and conclusions are different. For example, in the guest satisfaction survey of Sections 12.1 and 12.2, there is evidence of a significant difference between the hotels with respect to the proportion of guests who would return. From a different viewpoint, you could conclude that there is a significant relationship between the hotels and the likelihood that a guest would return. However, the two types of tests differ in how the samples are selected.

In a test for equality of proportions, there is one factor of interest, with two or more levels. These levels represent samples selected from independent populations. The categorical responses in each group or level are classified into two categories, such as *an item of interest* and *not an item of interest*. The objective is to make comparisons and evaluate differences between the proportions of the *items of interest* among the various levels. However, in a test for independence, there are two factors of interest, each of which has two or more levels. You select one sample and tally the joint responses to the two categorical variables into the cells of a contingency table.

To illustrate the χ^2 test for independence, suppose that, in the three-hotel guest satisfaction survey, respondents who stated that they were not likely to return also indicated the primary reason for their unwillingness to return. Table 12.9 presents the resulting 4×3 contingency table.

TABLE 12.9

Contingency Table of Primary Reason for Not Returning and Hotel

PRIMARY REASON FOR NOT RETURNING	HOTEL			Total
	Golden Palm	Palm Royale	Palm Princess	
Price	23	7	37	67
Location	39	13	8	60
Room accommodation	13	5	13	31
Other	13	8	8	29
Total	88	33	66	187

In Table 12.9, observe that of the primary reasons for not planning to return to the hotel, 67 were due to price, 60 were due to location, 31 were due to room accommodation, and 29 were due to other reasons. In Table 12.6 on page 456, there were 88 guests at the Golden Palm, 33 guests at the Palm Royale, and 66 guests at the Palm Princess who were not planning to return. The observed frequencies in the cells of the 4×3 contingency table represent the joint tallies of the sampled guests with respect to primary reason for not returning and the hotel where they stayed. The null and alternative hypotheses are

H_0 : There is no relationship between the primary reason for not returning and the hotel.

H_1 : There is a relationship between the primary reason for not returning and the hotel.

To test this null hypothesis of independence against the alternative that there is a relationship between the two categorical variables, you use Equation (12.1) on page 449 to compute the test statistic:

$$\chi^2_{STAT} = \sum_{all\ cells} \frac{(f_o - f_e)^2}{f_e}$$

where

f_o = observed frequency in a particular cell of the $r \times c$ contingency table

f_e = expected frequency in a particular cell if the null hypothesis of independence is true

To compute the expected frequency, f_e , in any cell, you use the multiplication rule for independent events discussed on page 166 [see Equation (4.7)]. For example, under the null hypothesis of independence, the probability of responses expected in the upper-left-corner cell representing primary reason of price for the Golden Palm is the product of the two separate probabilities $P(\text{Price})$ and $P(\text{Golden Palm})$. Here, the proportion of reasons that are due to price, $P(\text{Price})$, is $67/187 = 0.3583$, and the proportion of all responses from the Golden Palm, $P(\text{Golden Palm})$, is $88/187 = 0.4706$. If the null hypothesis is true, then the primary reason for not returning and the hotel are independent:

$$\begin{aligned} P(\text{Price and Golden Palm}) &= P(\text{Price}) \times P(\text{Golden Palm}) \\ &= (0.3583) \times (0.4706) \\ &= 0.1686 \end{aligned}$$

The expected frequency is the product of the overall sample size, n , and this probability, $187 \times 0.1686 = 31.53$. The f_e values for the remaining cells are shown in Table 12.10.

TABLE 12.10

Contingency Table of Expected Frequencies of Primary Reason for Not Returning with Hotel

PRIMARY REASON FOR NOT RETURNING	HOTEL			Total
	Golden Palm	Palm Royale	Palm Princess	
Price	31.53	11.82	23.65	67
Location	28.24	10.59	21.18	60
Room accommodation	14.59	5.47	10.94	31
Other	13.65	5.12	10.24	29
Total	88.00	33.00	66.00	187

You can also compute the expected frequency by taking the product of the row total and column total for a cell and dividing this product by the overall sample size, as Equation (12.5) shows.

COMPUTING THE EXPECTED FREQUENCY

The expected frequency in a cell is the product of its row total and column total, divided by the overall sample size.

$$f_e = \frac{\text{Row total} \times \text{Column total}}{n} \quad (12.5)$$

where

Row total = sum of the frequencies in the row

Column total = sum of the frequencies in the column

n = overall sample size

This alternate method results in simpler computations. For example, using Equation (12.5) for the upper-left-corner cell (price for the Golden Palm),

$$f_e = \frac{\text{Row total} \times \text{Column total}}{n} = \frac{(67)(88)}{187} = 31.53$$

and for the lower-right-corner cell (other reason for the Palm Princess),

$$f_e = \frac{\text{Row total} \times \text{Column total}}{n} = \frac{(29)(66)}{187} = 10.24$$

To perform the test of independence, you use the χ^2_{STAT} test statistic shown in Equation (12.1) on page 449. The sampling distribution of the χ^2_{STAT} test statistic approximately follows a chi-square distribution, with degrees of freedom equal to the number of rows in the contingency table minus 1, multiplied by the number of columns in the table minus 1:

$$\begin{aligned}\text{Degrees of freedom} &= (r - 1)(c - 1) \\ &= (4 - 1)(3 - 1) = 6\end{aligned}$$

Table 12.11 presents the computations for the χ^2_{STAT} test statistic.

TABLE 12.11

Computing the χ^2_{STAT} Test Statistic for the Test of Independence

Cell	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Price/Golden Palm	23	31.53	-8.53	72.76	2.31
Price/Palm Royale	7	11.82	-4.82	23.23	1.97
Price/Palm Princess	37	23.65	13.35	178.22	7.54
Location/Golden Palm	39	28.24	10.76	115.78	4.10
Location/Palm Royale	13	10.59	2.41	5.81	0.55
Location/Palm Princess	8	21.18	-13.18	173.71	8.20
Room/Golden Palm	13	14.59	-1.59	2.53	0.17
Room/Palm Royale	5	5.47	-0.47	0.22	0.04
Room/Palm Princess	13	10.94	2.06	4.24	0.39
Other/Golden Palm	13	13.65	-0.65	0.42	0.03
Other/Palm Royale	8	5.12	2.88	8.29	1.62
Other/Palm Princess	8	10.24	-2.24	5.02	0.49
					27.41

Using the $\alpha = 0.05$ level of significance, the upper-tail critical value from the chi-square distribution with 6 degrees of freedom is 12.592 (see Table E.4). Because $\chi^2_{STAT} = 27.41 > 12.592$, you reject the null hypothesis of independence (see Figure 12.9).

FIGURE 12.9

Regions of rejection and nonrejection when testing for independence in the three hotel guest satisfaction survey example at the 0.05 level of significance, with 6 degrees of freedom

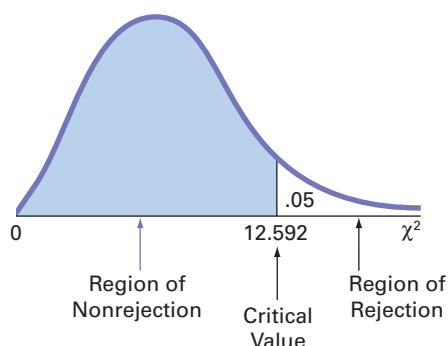


Figure 12.10 shows the Excel and Minitab results for this test, which are identical when rounded to three decimal places. Because $\chi^2_{STAT} = 27.410 > 12.592$, you reject the null hypothesis of independence. Using the p -value approach, you reject the null hypothesis of independence because the p -value = 0.000 < 0.05. The p -value indicates that there is virtually no chance of having a relationship this strong or stronger between the hotel and the primary reasons for not returning in a sample, if the primary reasons for not returning are independent of the specific hotels in the entire population. Thus, there is strong evidence of a relationship between the primary reason for not returning and the hotel.

FIGURE 12.10

Excel and Minitab chi-square test results for the Table 12.9 primary reason for not returning to hotel data

Chi-Square Test of Independence				
Observed Frequencies				
	Hotel			
Reason for Not Returning	Golden Palm	Palm Royale	Palm Princess	Total
Price	23	7	37	67
Location	39	13	8	60
Room accommodation	13	5	13	31
Other	13	8	8	29
Total	88	33	66	187
Expected Frequencies				
	Hotel			
Reason for Not Returning	Golden Palm	Palm Royale	Palm Princess	Total
Price	31.5294	11.8235	23.6471	67
Location	28.2353	10.5882	21.1765	60
Room accommodation	14.5882	5.4706	10.9412	31
Other	13.6471	5.1176	10.2353	29
Total	88	33	66	187
Data				
Level of Significance	0.05			
Number of Rows	4			
Number of Columns	3			
Degrees of Freedom	6	=B23 - 1) * (B24 - 1)		
Results				
Critical Value	12.5916	=CHISQ.INV.RT(B22, B25)		
Chi-Square Test Statistic	27.4104	=SUM(G15:I18)		
p-Value	0.0001	=CHISQ.DIST.RT(B29, B25)		
Reject the null hypothesis		=IF(B30 < B22, "Reject the null hypothesis", "Do not reject the null hypothesis")		
Expected frequency assumption is met.		=IF(OR(B15 < 1, C15 < 1, D15 < 1, B16 < 1, C16 < 1, D16 < 1, B17 < 1, C17 < 1, D17 < 1, B18 < 1, C18 < 1, D18 < 1), " is violated.", " is met.")		

Chi-Square Test: Golden Palm, Palm Royale, Palm Princess

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

	Golden	Palm	Palm	Total
1	23	7	37	67
	31.53	11.82	23.65	
	2.307	1.968	7.540	
2	39	13	8	60
	28.24	10.59	21.18	
	4.104	0.549	8.199	
3	13	5	13	31
	14.59	5.47	10.94	
	0.173	0.040	0.387	
4	13	8	8	29
	13.65	5.12	10.24	
	0.031	1.623	0.488	
Total	88	33	66	187

Chi-Sq = 27.410, DF = 6, P-Value = 0.000

Examination of the observed and expected frequencies (see Table 12.11 above on page 465) reveals that price is underrepresented as a reason for not returning to the Golden Palm (i.e., $f_o = 23$ and $f_e = 31.53$) but is overrepresented at the Palm Princess. Guests are more satisfied with the price at the Golden Palm than at the Palm Princess. Location is overrepresented as a reason for not returning to the Golden Palm but greatly underrepresented at the Palm Princess. Thus, guests are much more satisfied with the location of the Palm Princess than with that of the Golden Palm.

To ensure accurate results, all expected frequencies need to be large in order to use the χ^2 test when dealing with $r \times c$ contingency tables. As in the case of $2 \times c$ contingency tables in Section 12.2, all expected frequencies should be at least 1. For contingency tables in which one or more expected frequencies are less than 1, you can use the chi-square test after collapsing two or more low-frequency rows into one row (or collapsing two or more low-frequency columns into one column). Merging rows or columns usually results in expected frequencies sufficiently large to ensure the accuracy of the χ^2 test.

Problems for Section 12.3

LEARNING THE BASICS

12.20 If a contingency table has three rows and four columns, how many degrees of freedom are there for the χ^2 test of independence?

12.21 When performing a χ^2 test of independence in a contingency table with r rows and c columns, determine the upper-tail critical value of the test statistic in each of the following circumstances:

- $\alpha = 0.05$, $r = 4$ rows, $c = 5$ columns
- $\alpha = 0.01$, $r = 4$ rows, $c = 5$ columns
- $\alpha = 0.01$, $r = 4$ rows, $c = 6$ columns
- $\alpha = 0.01$, $r = 3$ rows, $c = 6$ columns
- $\alpha = 0.01$, $r = 6$ rows, $c = 3$ columns

APPLYING THE CONCEPTS

12.22 The owner of a restaurant serving Continental-style entrées has the business objective of learning more about the patterns of patron demand during the Friday-to-Sunday weekend time period. Data were collected from 630 customers on the type of entrée ordered and the type of dessert ordered and organized into the following table:

TYPE OF DESSERT	TYPE OF ENTRÉE				Total
	Beef	Poultry	Fish	Pasta	
Ice cream	13	8	12	14	47
Cake	98	12	29	6	145
Fruit	8	10	6	2	26
None	124	98	149	41	412
Total	243	128	196	63	630

At the 0.05 level of significance, is there evidence of a relationship between type of dessert and type of entrée?

12.23 Is there a generation gap in the type of music that people listen to? The following table represents the type of favorite music for a sample of 1,000 respondents classified according to their age group:

FAVORITE TYPE	AGE				Total
	16–29	30–49	50–64	65 and over	
Rock	71	62	51	27	211
Rap or hip-hop	40	21	7	3	71
Rhythm and blues	48	46	46	40	180
Country	43	53	59	79	234
Classical	22	28	33	46	129
Jazz	18	26	36	43	123
Salsa	8	14	18	12	52
Total	250	250	250	250	1000

At the 0.05 level of significance, is there evidence of a relationship between favorite type of music and age group?

 **12.24** How often do Facebook users post? A study by the Pew Research Center (data extracted from bit.ly/1axbobZ) revealed the following results:

FREQUENCY	AGE GROUP					Total
	18–22	23–35	36–49	50–65	65+	
Several times a day	20	22	9	2	1	54
About once a day	28	37	14	4	1	84
3–5 days per week	32	46	31	7	2	118
1–2 days per week	32	69	35	17	7	160
Every few weeks	22	66	47	28	6	169
Less often	16	59	56	61	18	210
Never	6	15	42	66	23	152
Total	156	314	234	185	58	947

At the 0.01 level of significance, is there evidence of a significant relationship between frequency of posting on Facebook and age?

12.25 Where people look for news is different for various age groups. A study indicated where different age groups primarily get their news:

MEDIA	AGE GROUP		
	Under 36	36–50	50 +
Local TV	107	119	133
National TV	73	102	127
Radio	75	97	109
Local newspaper	52	79	107
Internet	95	83	76

At the 0.05 level of significance, is there evidence of a significant relationship between the age group and where people primarily get their news? If so, explain the relationship.

12.26 The 2012 Restaurant Industry Forecast takes a closer look at today's consumers. Based on a 2011 National Restaurant Association survey, American adults are categorized into one of

three consumer segments (optimistic, cautious, and hunkered down) based on their financial situation, current spending behavior, and economic outlook, as well as the geographic region where they reside. Suppose the results, based on a sample 1,000 American adults, are as follows:

CONSUMER SEGMENT	GEOGRAPHIC REGION				Total
	Midwest	Northeast	South	West	
Optimistic	67	23	60	63	213
Cautious	101	57	127	133	418
Hunkered down	83	46	115	125	369
Total	251	126	302	321	1000

Source: Data extracted from "The 2012 Restaurant Industry Forecast," National Restaurant Association, 2012, p. 12.

At the 0.05 level of significance, is there evidence of a significant relationship between consumer segment and geographic region?

12.4 Wilcoxon Rank Sum Test: A Nonparametric Method for Two Independent Populations

In Section 10.1, you used the *t* test for the difference between the means of two independent populations. If sample sizes are small and you cannot assume that the data in each sample are from normally distributed populations, you have two choices:

- Use a *nonparametric* method that does not depend on the assumption of normality for the two populations.
- Use the pooled-variance *t* test, following a *normalizing transformation* on the data (see reference 10).

Nonparametric methods require few or no assumptions about the populations from which data are obtained (see reference 4). One such method is the Wilcoxon rank sum test for testing whether there is a difference between two *medians*. The **Wilcoxon rank sum test** is almost as powerful as the pooled-variance and separate-variance *t* tests discussed in Section 10.1 under conditions appropriate to these tests and is likely to be more powerful when the assumptions of those *t* tests are not met. In addition, you can use the Wilcoxon rank sum test when you have only ordinal data, as often happens in consumer behavior and marketing research.

To perform the Wilcoxon rank sum test, you replace the values in the two samples of sizes n_1 and n_2 with their combined ranks (unless the data contained the ranks initially). You begin by defining $n = n_1 + n_2$ as the total sample size. Next, you assign the ranks so that rank 1 is given to the smallest of the n combined values, rank 2 is given to the second smallest, and so on, until rank n is given to the largest. If several values are tied, you assign each value the average of the ranks that otherwise would have been assigned had there been no ties.

Whenever the two sample sizes are unequal, n_1 represents the smaller sample and n_2 the larger sample. The Wilcoxon rank sum test statistic, T_1 , is defined as the sum of the ranks assigned to the n_1 values in the smaller sample. (For equal-sized samples, either sample may be used for determining T_1 .) For any integer value n , the sum of the first n consecutive integers is $n(n + 1)/2$. Therefore, T_1 plus T_2 , the sum of the ranks assigned to the n_2 items in the second sample, must equal $n(n + 1)/2$. You can use Equation (12.6) to check the accuracy of your rankings.

Student Tip

Remember that you combine the two groups before you rank the values.

CHECKING THE RANKINGS

$$T_1 + T_2 = \frac{n(n + 1)}{2} \quad (12.6)$$

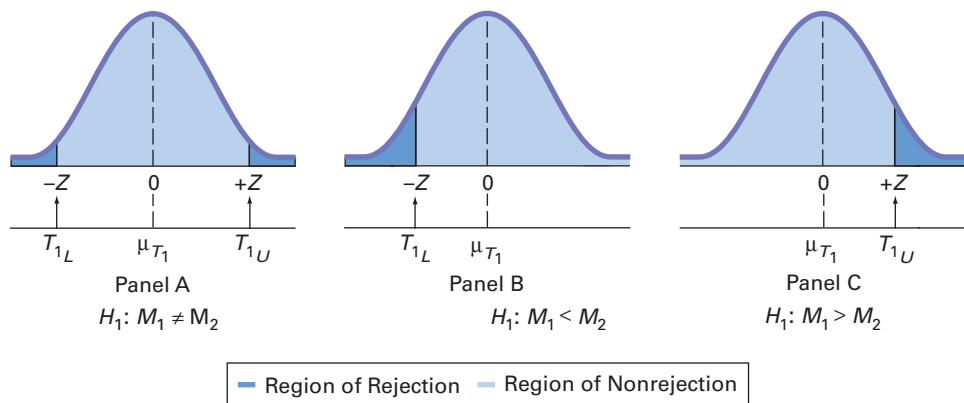
When n_1 and n_2 are each ≤ 10 , you use Table E.6 to find the critical values of the test statistic T_1 . For a two-tail test shown in Figure 12.11 Panel A, you reject the null hypothesis if the computed value of T_1 is greater than or equal to the upper critical value, or if T_1 is less than or equal to the lower critical value. For one-tail tests having the alternative hypothesis $H_1: M_1 < M_2$ that the median of population 1 (M_1) is less than the median of population 2 (M_2), you reject the null hypothesis if the observed value of T_1 is less than or equal to the lower critical value (shown in Figure 12.11 Panel B). For one-tail tests having the alternative hypothesis $H_1: M_1 > M_2$, you reject the null hypothesis if the observed value of T_1 equals or is greater than the upper critical value (shown in Figure 12.11 Panel C).

FIGURE 12.11

Regions of rejection and nonrejection using the Wilcoxon rank sum test

Student Tip

Remember that the group that is defined as group 1 when computing the test statistic T_1 must also be defined as group 1 in the null and alternative hypotheses.



For large sample sizes, the test statistic T_1 is approximately normally distributed, with the mean, μ_{T_1} , equal to

$$\mu_{T_1} = \frac{n_1(n + 1)}{2}$$

and the standard deviation, σ_{T_1} , equal to

$$\sigma_{T_1} = \sqrt{\frac{n_1 n_2 (n + 1)}{12}}$$

Therefore, Equation (12.7) defines the standardized Z test statistic for the Wilcoxon rank sum test.

LARGE-SAMPLE WILCOXON RANK SUM TEST

$$Z_{STAT} = \frac{T_1 - \frac{n_1(n + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n + 1)}{12}}} \quad (12.7)$$

where the test statistic Z_{STAT} approximately follows a standardized normal distribution.

You use Equation (12.7) when the sample sizes are outside the range of Table E.6. Based on α , the level of significance selected, you reject the null hypothesis if the Z_{STAT} test statistic falls in the rejection region.

²To test for differences in the median sales between the two locations, you must assume that the distributions of sales in both populations are identical except for differences in central tendency (i.e., the medians).

To study an application of the Wilcoxon rank sum test, recall the Chapter 10 Using Statistics scenario concerning cola sales for two different end-cap display locations (stored in **Cola**). If you cannot assume that the populations are normally distributed, you can use the Wilcoxon rank sum test to evaluate possible differences in the median sales for the two display locations.² The cola sales data and the combined ranks are shown in Table 12.12.

TABLE 12.12

Forming the Combined Rankings

Beverage End-cap ($n_1 = 10$)	Combined Ranking	Produce End-cap ($n_2 = 10$)	Combined Ranking
22	1.0	52	5.5
34	3.0	71	14.0
52	5.5	76	15.0
62	10.0	54	7.0
30	2.0	67	13.0
40	4.0	83	17.0
64	11.0	66	12.0
84	18.5	90	20.0
56	8.0	77	16.0
59	9.0	84	18.5

Source: Data are taken from Table 10.1 on page 349.

Because you have not stated in advance which display location is likely to have a higher median, you use a two-tail test with the following null and alternative hypotheses:

$$H_0: M_1 = M_2 \text{ (the median sales are equal)}$$

$$H_1: M_1 \neq M_2 \text{ (the median sales are not equal)}$$

Next, you compute T_1 , the sum of the ranks assigned to the *smaller* sample. When the sample sizes are equal, as in this example, you can define either sample as the group from which to compute T_1 . Choosing the beverage end-cap display as the first group,

$$T_1 = 1 + 3 + 5.5 + 10 + 2 + 4 + 11 + 18.5 + 8 + 9 = 72$$

As a check on the ranking procedure, you compute T_2 from

$$T_2 = 5.5 + 14 + 15 + 7 + 13 + 17 + 12 + 20 + 16 + 18.5 = 138$$

and then use Equation (12.6) on page 468 to show that the sum of the first $n = 20$ integers in the combined ranking is equal to $T_1 + T_2$:

$$T_1 + T_2 = \frac{n(n + 1)}{2}$$

$$72 + 138 = \frac{20(21)}{2} = 210$$

$$210 = 210$$

Next, you use Table E.6 to determine the lower- and upper-tail critical values for the test statistic T_1 . From Table 12.13, a portion of Table E.6, observe that for a level of significance of 0.05, the critical values are 78 and 132. The decision rule is

Reject H_0 if $T_1 \leq 78$ or if $T_1 \geq 132$;

otherwise, do not reject H_0 .

TABLE 12.13

Finding the Lower-
and Upper-Tail
Critical Values for the
Wilcoxon Rank Sum
Test Statistic, T_1 , Where
 $n_1 = 10$, $n_2 = 10$, and
 $\alpha = 0.05$

n_2	α	n_1								10
		One-tail	Two-tail	4	5	6	7	(Lower, Upper)	9	
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: Extracted from Table E.6.

Because the test statistic $T_1 = 72 < 78$, you reject H_0 . There is evidence of a significant difference in the median sales for the two display locations. Because the sum of the ranks is lower for the beverage end-cap display, you conclude that median sales are lower for the beverage end-cap display.

Figure 12.12 shows the Wilcoxon rank sum test results (Excel) and the equivalent Mann-Whitney test results (Minitab) for the cola sales data. The p -values differ because Minitab adjusts for ties and computes an exact probability while Excel uses the normal approximation. From these results, you reject the null hypothesis because the p -value is 0.0126 (0.0139 in Minitab results), which is less than $\alpha = 0.05$. This p -value indicates that if the medians of the two populations are equal, the chance of finding a difference at least this large in the samples is only 0.0126 (0.0139 in Minitab).

FIGURE 12.12

Wilcoxon rank sum test and Mann-Whitney test (Minitab) results for cola sales for two different end-cap locations

Wilcoxon Rank Sum Test		Mann-Whitney Test and CI: Beverage, Produce	
Data		N Median	
Level of Significance		Beverage	10 54.00
Population 1 Sample		Produce	10 73.50
Sample Size	10	Point estimate for ETA1-ETA2 is -21.50	
Sum of Ranks	72	95.5 Percent CI for ETA1-ETA2 is (-37.01, -6.00)	
Population 2 Sample		W = 72.0	
Sample Size	10	Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0140	
Sum of Ranks	138	The test is significant at 0.0139 (adjusted for ties)	
Intermediate Calculations			
Total Sample Size	20		
T_1 : Test Statistic	72		
T_1 Mean	105		
Standard Error of T_1	13.2288		
Z Test Statistic	-2.4946		
Two-Tail Test			
Lower Critical Value	-1.9600		
Upper Critical Value	1.9600		
p-Value	0.0126		
Reject the null hypothesis	"Do not reject the null hypothesis"		

Table E.6 shows the lower and upper critical values of the Wilcoxon rank sum test statistic, T_1 , but only for situations in which both n_1 and n_2 are less than or equal to 10. If either one or both of the sample sizes are greater than 10, you *must* use the large-sample Z approximation formula [Equation (12.7) on page 468]. However, you can also use this approximation formula

for small sample sizes. To demonstrate the large-sample Z approximation formula, consider the cola sales data. Using Equation (12.7),

$$\begin{aligned} Z_{STAT} &= \frac{T_1 - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_2(n+1)}{12}}} \\ &= \frac{72 - \frac{(10)(21)}{2}}{\sqrt{\frac{(10)(10)(21)}{12}}} \\ &= \frac{72 - 105}{13.2288} = -2.4946 \end{aligned}$$

Because $Z_{STAT} = -2.4946 < -1.96$, the critical value of Z at the 0.05 level of significance (or p -value = 0.0126 < 0.05), you reject H_0 .

Problems for Section 12.4

LEARNING THE BASICS

12.27 Using Table E.6, determine the lower- and upper-tail critical values for the Wilcoxon rank sum test statistic, T_1 , in each of the following two-tail tests:

- a. $\alpha = 0.10, n_1 = 6, n_2 = 8$
- b. $\alpha = 0.05, n_1 = 6, n_2 = 8$
- c. $\alpha = 0.01, n_1 = 6, n_2 = 8$
- d. Given the results in (a) through (c), what do you conclude regarding the width of the region of nonrejection as the selected level of significance, α , gets smaller?

12.28 Using Table E.6, determine the lower-tail critical value for the Wilcoxon rank sum test statistic, T_1 , in each of the following one-tail tests:

- a. $\alpha = 0.05, n_1 = 6, n_2 = 8$
- b. $\alpha = 0.025, n_1 = 6, n_2 = 8$
- c. $\alpha = 0.01, n_1 = 6, n_2 = 8$
- d. $\alpha = 0.005, n_1 = 6, n_2 = 8$

12.29 The following information is available for two samples selected from independent populations:

Sample 1: $n_1 = 7$ Assigned ranks: 4 1 8 2 5 10 11

Sample 2: $n_2 = 9$ Assigned ranks: 7 16 12 9 3 14 13 6 15

What is the value of T_1 if you are testing the null hypothesis $H_0: M_1 = M_2$?

12.30 In Problem 12.29, what are the lower- and upper-tail critical values for the test statistic T_1 from Table E.6 if you use a 0.05 level of significance and the alternative hypothesis is $H_1: M_1 \neq M_2$?

12.31 In Problems 12.29 and 12.30, what is your statistical decision?

12.32 The following information is available for two samples selected from independent and similarly shaped right-skewed populations:

Sample 1: $n_1 = 5$ 1.1 2.3 2.9 3.6 14.7

Sample 2: $n_2 = 6$ 2.8 4.4 4.4 5.2 6.0 18.5

- a. Replace the observed values with the corresponding ranks (where 1 = smallest value; $n = n_1 + n_2 = 11$ = largest value) in the combined samples.
- b. What is the value of the test statistic T_1 ?
- c. Compute the value of T_2 , the sum of the ranks in the larger sample.
- d. To check the accuracy of your rankings, use Equation (12.6) on page 468 to demonstrate that $T_1 + T_2 = \frac{n(n+1)}{2}$

12.33 From Problem 12.32, at the 0.05 level of significance, determine the lower-tail critical value for the Wilcoxon rank sum test statistic, T_1 , if you want to test the null hypothesis, $H_0: M_1 \geq M_2$, against the one-tail alternative, $H_1: M_1 < M_2$.

12.34 In Problems 12.32 and 12.33, what is your statistical decision?

APPLYING THE CONCEPTS

12.35 A vice president for marketing recruits 20 college graduates for management training. Each of the 20 individuals is randomly assigned to one of two groups (10 in each group). A “traditional” method of training (T) is used in one group, and an “experimental” method (E) is used in the other. After the graduates spend six months on the job, the vice president ranks them on the basis of their performance, from 1 (worst) to 20 (best), with the following results (stored in the file **TestRank**):

T: 1 2 3 5 9 10 12 13 14 15

E: 4 6 7 8 11 16 17 18 19 20

Is there evidence of a difference in the median performance between the two methods? (Use $\alpha = 0.05$.)

12.36 Wine experts Gaiter and Brecher use a six-category scale when rating wines: Yech, OK, Good, Very Good, Delicious, and Delicious! Suppose Gaiter and Brecher tested wines from a random sample of eight inexpensive California Cabernets and a random sample of eight inexpensive Washington Cabernets, where *inexpensive* means wines with a U.S. suggested retail price of less than \$20, and assigned the following ratings:

California—Good, Delicious, Yech, OK, OK, Very Good, Yech, OK

Washington—Very Good, OK, Delicious!, Very Good, Delicious, Good, Delicious, Delicious!

The ratings were then ranked and the ratings and the rankings stored in **Cabernet**. (Data extracted from D. Gaiter and J. Brecher, “A Good U.S. Cabernet Is Hard to Find,” *The Wall Street Journal*, May 19, 2006, p. W7.)

- Are the data collected by rating wines using this scale nominal, ordinal, interval, or ratio?
- Why is the two-sample *t* test defined in Section 10.1 inappropriate to test the mean rating of California Cabernets versus Washington Cabernets?
- Is there evidence of a significant difference in the median rating of California Cabernets and Washington Cabernets? (Use $\alpha = 0.05$.)

12.37 A problem with a telephone line that prevents a customer from receiving or making calls is upsetting to both the customer and the telephone company. The file **Phone** contains samples of 20 problems reported to two different offices of a telecommunications company and the time to clear these problems (in minutes) from the customers’ lines:

Central Office I Time to Clear Problems (Minutes)

1.48	1.75	0.78	2.85	0.52	1.60	4.15	3.97	1.48	3.10
1.02	0.53	0.93	1.60	0.80	1.05	6.32	3.93	5.45	0.97

Central Office II Time to Clear Problems (Minutes)

7.55	3.75	0.10	1.10	0.60	0.52	3.30	2.10	0.58	4.02
3.75	0.65	1.92	0.60	1.53	4.23	0.08	1.48	1.65	0.72

- Is there evidence of a difference in the median time to clear problems between offices? (Use $\alpha = 0.05$.)
- What assumptions must you make in (a)?
- Compare the results of (a) with those of Problem 10.9(a) on page 357.

 **SELF Test** **12.38** The management of a hotel has the business objective of increasing the return rate for hotel guests. One aspect of first impressions by guests relates to the time it takes to deliver a guest’s luggage to the room after check-in to the hotel. A random sample of 20 deliveries on a particular day were selected in Wing A of the hotel, and a random sample of 20 deliveries were selected in Wing B. Delivery times were collected and stored in **Luggage**.

- Is there evidence of a difference in the median delivery times in the two wings of the hotel? (Use $\alpha = 0.05$.)
- Compare the results of (a) with those of Problem 10.65 on page 382.

12.39 The lengths of life (in hours) of a sample of 40 100-watt compact fluorescent light bulbs produced by Manufacturer A and

a sample of 40 20-watt compact fluorescent light bulbs produced by Manufacturer B are stored in **Bulbs**.

- Using a 0.05 level of significance, is there evidence of a difference in the median life of bulbs produced by the two manufacturers?
- What assumptions must you make in (a)?
- Compare the results of (a) with those of Problem 10.64 on page 382. Discuss.

12.40 Brand valuations are critical to CEOs, financial and marketing executives, security analysts, institutional investors, and others who depend on well-researched, reliable information for assessments and comparisons in decision making. Millward Brown, Inc., has annually compiled its BrandZ Top 100 Most Valuable Global Brands since 1996. Unlike other studies, the BrandZ rankings combines consumer measures of brand equity with financial measures to establish a *brand value* for each brands. The file **BrandZTechFin** contains the brand values for two sectors in the BrandZ Top 100 Most Valuable Global Brands for 2013: the technology sector and the financial institutions sector. (Data extracted from “BrandZ Top 1000 Most Valuable Global Brands 2011,” Millward Brown, Inc., retrieved from bit.ly/18OL5Mu.)

- Using a 0.05 level of significance, is there evidence of a difference in the median brand value between the two sectors?
- What assumptions must you make in (a)?
- Compare the results of (a) with those of Problem 10.17 on page 358. Discuss.

12.41 A bank with a branch located in a commercial district of a city has developed an improved process for serving customers during the noon-to-1 P.M. lunch period. The bank has the business objective of reducing the waiting time (defined as the number of minutes that elapse from when the customer enters the line until he or she reaches the teller window) to increase customer satisfaction. A random sample of 15 customers is selected and waiting times are collected and stored in **Bank1**. These waiting times (in minutes) are:

4.21	5.55	3.02	5.13	4.77	2.34	3.54	3.20
4.50	6.10	0.38	5.12	6.46	6.19	3.79	

Another branch, located in a residential area, is also concerned with the noon-to-1 P.M. lunch period. A random sample of 15 customers is selected and waiting times are collected and stored in **Bank2**. These waiting times (in minutes) are:

9.66	5.90	8.02	5.79	8.73	3.82	8.01	8.35
10.49	6.68	5.64	4.08	6.17	9.91	5.47	

- Is there evidence of a difference in the median waiting time between the two branches? (Use $\alpha = 0.05$.)
- What assumptions must you make in (a)?
- Compare the results (a) with those of Problem 10.12 (a) on page 358. Discuss.

12.42 An important feature of digital cameras is battery life, the number of shots that can be taken before the battery needs to be recharged. The file **Cameras** contains the battery life of 11 subcompact cameras and 7 compact cameras. (Data extracted from “Cameras,” *Consumer Reports*, July 2012, pp. 42–44.)

- Is there evidence of a difference in the median battery life between subcompact cameras and compact cameras? (Use $\alpha = 0.05$.)
- What assumptions must you make in (a)?
- Compare the results of (a) with those of Problem 10.11 (a) on page 358. Discuss.

12.5 Kruskal-Wallis Rank Test: A Nonparametric Method for the One-Way ANOVA

If the normality assumption of the one-way ANOVA F test is violated, you can use the Kruskal-Wallis rank test. The **Kruskal-Wallis rank test** for differences among c medians (where $c > 2$) is an extension of the Wilcoxon rank sum test for two independent populations, discussed in Section 12.4. The Kruskal-Wallis test has the same power relative to the one-way ANOVA F test that the Wilcoxon rank sum test has relative to the t test.

You use the Kruskal-Wallis rank test to test whether c independent groups have equal medians. The null hypothesis is

$$H_0: M_1 = M_2 = \dots = M_c$$

and the alternative hypothesis is

$$H_1: \text{Not all } M_j \text{ are equal (where } j = 1, 2, \dots, c\text{).}$$

To use the Kruskal-Wallis rank test, you first replace the values in the c samples with their combined ranks (if necessary). Rank 1 is given to the smallest of the combined values and rank n to the largest of the combined values (where $n = n_1 + n_2 + \dots + n_c$). If any values are tied, you assign each of them the mean of the ranks they would have otherwise been assigned if ties had not been present in the data.

The Kruskal-Wallis test is an alternative to the one-way ANOVA F test. Instead of comparing each of the c group means against the grand mean, the Kruskal-Wallis test compares the mean rank in each of the c groups against the overall mean rank, based on all n combined values. Equation (12.8) defines the Kruskal-Wallis test statistic, H .

Student Tip

Remember that you combine the groups before you rank the values.

KRUSKAL-WALLIS RANK TEST FOR DIFFERENCES AMONG c MEDIANs

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^c \frac{T_j^2}{n_j} \right] - 3(n+1) \quad (12.8)$$

where

n = total number of values over the combined samples

n_j = number of values in the j th sample ($j = 1, 2, \dots, c$)

T_j = sum of the ranks assigned to the j th sample

T_j^2 = square of the sum of the ranks assigned to the j th sample

c = number of groups

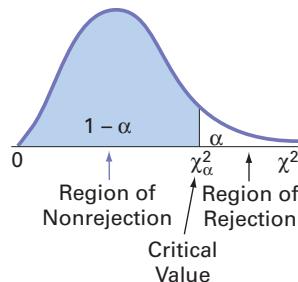
If there is a significant difference among the c groups, the mean rank differs considerably from group to group. In the process of squaring these differences, the test statistic H becomes large. If there are no differences present, the test statistic H is small because the mean of the ranks assigned in each group should be very similar from group to group.

As the sample sizes in each group get large (i.e., at least 5), the sampling distribution of the test statistic, H , approximately follows the chi-square distribution with $c - 1$ degrees of freedom. Thus, you reject the null hypothesis if the computed value of H is greater than the upper-tail critical value (see Figure 12.13). Therefore, the decision rule is

$$\begin{aligned} \text{Reject } H_0 \text{ if } H > \chi_{\alpha}^2; \\ \text{otherwise, do not reject } H_0. \end{aligned}$$

FIGURE 12.13

Determining the rejection region for the Kruskal-Wallis test



To illustrate the Kruskal-Wallis rank test for differences among c medians, return to the Arlington's scenario on page 394 that concerns the in-store sales location experiment. If you cannot assume that the mobile electronics sales is normally distributed in all c groups, you can use the Kruskal-Wallis rank test.

The null hypothesis is that the median mobile electronics sales from each of the four in-store locations are equal. The alternative hypothesis is that at least one of these medians differs from the others:

$$H_0: M_1 = M_2 = M_3 = M_4$$

$$H_1: \text{Not all } M_j \text{ are equal (where } j = 1, 2, 3, 4\text{).}$$

Table 12.14 presents the data (stored in **Second Experiment**), along with the corresponding ranks of a second in-store location sales experiment at Arlington's.

TABLE 12.14

Mobile Electronics Sales and Rank for Four In-Store Locations for Second Experiment

IN-AISLE		FRONT		KIOSK		EXPERT	
Sales	Rank	Sales	Rank	Sales	Rank	Sales	Rank
29.60	2	33.21	20	31.23	14	29.83	6.5
29.59	1	31.80	17	31.14	13	29.69	4
30.11	9	30.54	11	30.04	8	29.61	3
29.83	6.5	31.39	15	30.87	12	29.72	5
30.32	10	33.05	18	33.08	19	31.41	16

In assigning ranks to the sales, the lowest sales, the second in-aisle sales in Table 12.4, is assigned the rank of 1 and the highest sales, the first front sales, is assigned the rank of 20. Because the fourth in-aisle sales and the first expert sales are tied for ranks 6 and 7, each is assigned the rank 6.5.

After all the ranks are assigned, you compute the sum of the ranks for each group:

$$\text{Rank sums: } T_1 = 28.5 \quad T_2 = 81 \quad T_3 = 66 \quad T_4 = 34.5$$

As a check on the rankings, recall from Equation (12.6) on page 468 that for any integer n , the sum of the first n consecutive integers is $n(n + 1)/2$. Therefore,

$$T_1 + T_2 + T_3 + T_4 = \frac{n(n + 1)}{2}$$

$$28.5 + 81 + 66 + 34.5 = \frac{(20)(21)}{2}$$

$$210 = 210$$

To test the null hypothesis of equal population medians, you calculate the test statistic H using Equation (12.8) on page 473:

$$\begin{aligned} H &= \left[\frac{12}{n(n+1)} \sum_{j=1}^c \frac{T_j^2}{n_j} \right] - 3(n+1) \\ &= \left\{ \frac{12}{(20)(21)} \left[\frac{(28.5)^2}{5} + \frac{(81)^2}{5} + \frac{(66)^2}{5} + \frac{(34.5)^2}{5} \right] \right\} - 3(21) \\ &= \left(\frac{12}{420} \right) (2,583.9) - 63 = 10.8257 \end{aligned}$$

The test statistic H approximately follows a chi-square distribution with $c - 1$ degrees of freedom. Using a 0.05 level of significance, χ_{α}^2 , the upper-tail critical value of the chi-square distribution with $c - 1 = 3$ degrees of freedom, is 7.815 (see Table 12.15). Because the computed value of the test statistic $H = 10.8257$ is greater than the critical value of 7.815, you reject the null hypothesis and conclude that the median mobile electronics sales is not the same for all the in-store locations.

TABLE 12.15

Finding χ_{α}^2 , the Upper-Tail Critical Value for the Kruskal-Wallis Rank Test, at the 0.05 Level of Significance with 3 Degrees of Freedom

Degrees of Freedom	Cumulative Area								
	.005	.01	.025	.05	.10	.25	.75	.90	.95
	Upper-Tail Area								
1	—	—	0.001	0.004	0.016	0.102	1.323	2.706	3.841
2	0.010	0.020	0.051	0.103	0.211	0.575	2.773	4.605	5.991
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.071
									12.833

Source: Extracted from Table E.4.

Figure 12.14 show the Excel and Minitab Kruskal-Wallis rank test results, which are identical when rounded to three decimal places. From these results, you reject the null hypothesis because the p -value = 0.0127 < 0.05. At this point, you could simultaneously compare all pairs of suppliers to determine which ones differ (see reference 2).

FIGURE 12.14

Excel and Minitab Kruskal-Wallis rank test results for the differences among the median mobile electronics sales for four in-store locations

Kruskal-Wallis Test: Sales versus Location					
Kruskal-Wallis Test on Sales					
A	B	C	D	E	G
1 Kruskal-Wallis Rank Test					
2					
3 Data					
4 Level of Significance	0.05				
5					
6 Intermediate Calculations					
7 Sum of Squared Ranks/Sample Size	2583.9				
8 Sum of Sample Sizes	20				
9 Number of Groups	4				
10					
11 Test Result					
12 H Test Statistic	10.8257	={12/(B8 * (B8+1))} * B7 - (3 * (B8+1))			
13 Critical Value	7.8147	=CHISQ.INV.RT(B4, B9 - 1)			
14 p-Value	0.0127	=CHISQ.DIST.RT(B12, B9 - 1)			
15 Reject the null hypothesis		=IF(B14 < B4, "Reject the null hypothesis", "Do not reject the null hypothesis")			
Also					
Cell B7: = (G5 * F5) + (G6 * F6) + (G7 * F7) + (G8 * F8)					
Cell B8: =SUM(E5:E8)					

Assumptions

To use the Kruskal-Wallis rank test, the following assumptions must be met:

- The c samples are randomly and independently selected from their respective populations.
- The underlying variable is continuous.
- The data provide at least a set of ranks, both within and among the c samples.
- The c populations have the same variability.
- The c populations have the same shape.

The Kruskal-Wallis procedure makes less stringent assumptions than does the F test. If you ignore the last two assumptions (variability and shape), you can still use the Kruskal-Wallis rank test to determine whether at least one of the populations differs from the other populations in some characteristic—such as central tendency, variation, or shape.

To use the F test, you must assume that the c samples are from normal populations that have equal variances. When the more stringent assumptions of the F test hold, you should use the F test instead of the Kruskal-Wallis test because it has slightly more power to detect significant differences among groups. However, if the assumptions of the F test do not hold, you should use the Kruskal-Wallis test.

Problems for Section 12.5

LEARNING THE BASICS

12.43 What is the upper-tail critical value from the chi-square distribution if you use the Kruskal-Wallis rank test for comparing the medians in six populations at the 0.01 level of significance?

12.44 For this problem, use the results of Problem 12.43.

- State the decision rule for testing the null hypothesis that all six groups have equal population medians.
- What is your statistical decision if the computed value of the test statistic H is 13.77?

APPLYING THE CONCEPTS

12.45 A pet food company has the business objective of expanding its product line beyond its current kidney- and shrimp-based cat foods. The company developed two new products—one based on chicken livers and the other based on salmon. The company conducted an experiment to compare the two new products with its two existing ones, as well as a generic beef-based product sold in a supermarket chain.

For the experiment, a sample of 50 cats from the population at a local animal shelter was selected. Ten cats were randomly assigned to each of the five products being tested. Each of the cats was then presented with 3 ounces of the selected food in a dish at feeding time. The researchers defined the variable to be measured as the number of ounces of food that the cat consumed within a 10-minute time interval that began when the filled dish was presented. The results for this experiment are summarized in the table at right and stored in **CatFood**.

- At the 0.05 level of significance, is there evidence of a significant difference in the median amount of food eaten among the various products?
- Compare the results of (a) with those of Problem 11.13 (a) on page 409.
- Which test is more appropriate for these data: the Kruskal-Wallis rank test or the one-way ANOVA F test? Explain.

Kidney	Shrimp	Chicken Liver	Salmon	Beef
2.37	2.26	2.29	1.79	2.09
2.62	2.69	2.23	2.33	1.87
2.31	2.25	2.41	1.96	1.67
2.47	2.45	2.68	2.05	1.64
2.59	2.34	2.25	2.26	2.16
2.62	2.37	2.17	2.24	1.75
2.34	2.22	2.37	1.96	1.18
2.47	2.56	2.26	1.58	1.92
2.45	2.36	2.45	2.18	1.32
2.32	2.59	2.57	1.93	1.94



12.46 A hospital conducted a study of the waiting time in its emergency room. The hospital has a main campus, along with three satellite locations. Management had a business objective of reducing waiting time for emergency room cases that did not require immediate attention. To study this, a random sample of 15 emergency room cases at each location were selected on a particular day, and the waiting time (recorded from check-in to when the patient was called into the clinic area) was measured. The results are stored in **ERWaiting**.

- At the 0.05 level of significance, is there evidence of a difference in the median waiting times in the four locations?
- Compare the results of (a) with those of Problem 11.9 (a) on page 408.

12.47 *QSR* magazine has been reporting on the largest quick-serve and fast-casual brands in the United States for nearly 15 years. The file **QSR** contains the food segment (burger, chicken, pizza, or sandwich) and U.S. mean sales per unit (\$thousands) for each of 38 quick-service brands. (Data extracted from bit.ly/Oj6EcY.)

- At the 0.05 level of significance, is there evidence of a difference in the median U.S. average sales per unit (\$thousands) among the food segments?
- Compare the results of (a) with those of Problem 11.11 (a) on page 408.

12.48 An advertising agency has been hired by a manufacturer of pens to develop an advertising campaign for the upcoming holiday season. To prepare for this project, the research director decides to initiate a study of the effect of advertising on product perception. An experiment is designed to compare five different advertisements. Advertisement A greatly undersells the pen's characteristics. Advertisement B slightly undersells the pen's characteristics. Advertisement C slightly oversells the pen's characteristics. Advertisement D greatly oversells the pen's characteristics. Advertisement E attempts to correctly state the pen's characteristics.

A sample of 30 adult respondents, taken from a larger focus group, is randomly assigned to the five advertisements (so that there are six respondents to each). After reading the advertisement and developing a sense of product expectation, all respondents unknowingly receive the same pen to evaluate. The respondents are permitted to test the pen and the plausibility of the advertising copy. The respondents are then asked to rate the pen from 1 to 7 on the product characteristic scales of appearance, durability, and writing performance. The *combined* scores of three ratings (appearance, durability, and writing performance) for the 30 respondents are stored in **Pen**. These data are:

A	B	C	D	E
15	16	8	5	12
18	17	7	6	19
17	21	10	13	18
19	16	15	11	12
19	19	14	9	17
20	17	14	10	14

- At the 0.05 level of significance, is there evidence of a difference in the median ratings of the five advertisements?
- Compare the results of (a) with those of Problem 11.10 (a) on page 408.
- Which test is more appropriate for these data: the Kruskal-Wallis rank test or the one-way ANOVA *F* test? Explain.

12.49 A sporting goods manufacturing company wanted to compare the distance traveled by golf balls produced using each of four different designs. Ten balls of each design were manufactured and brought to the local golf course for the club professional to test. The order in which the balls were hit with the same club from the first tee was randomized so that the pro did not know which type of ball was being hit. All 40 balls were hit in a short period of time, during which the environmental conditions were essentially the same. The results (distance traveled in yards) for the four designs are stored in **Golfball**.

- At the 0.05 level of significance, is there evidence of a difference in the median distances traveled by the golf balls with different designs?
- Compare the results of (a) with those of Problem 11.14 (a) on page 409.

12.50 The more costly and time consuming it is to export and import, the more difficult it is for local companies to be competitive and to reach international markets. As part of an initial investigation exploring foreign market entry, 10 countries were selected from each of four global regions. The cost associated with importing a standardized cargo of goods by sea transport in these countries (in US\$ per container) is stored in **ForeignMarket2**. (Data extracted from doingbusiness.org/data).

- At the 0.05 level of significance, is there evidence of a difference in the median cost across the four global regions associated with importing a standardized cargo of goods by sea transport?
- Compare the results in (a) to those in Problem 11.8 (a) on page 408.

12.6 McNemar Test for the Difference Between Two Proportions (Related Samples)

Tests such as chi-square test for the difference between two proportions discussed in Section 12.1 require independent samples from each population. However, sometimes when you are testing differences between the proportion of items of interest, the data are collected from repeated measurements or matched samples.

To test whether there is evidence of a difference between the proportions when the data have been collected from two related samples, you can use the McNemar test. The **Section 12.6 online topic** discusses this test and illustrates its use.

12.7 Chi-Square Test for the Variance or Standard Deviation

When analyzing numerical data, sometimes you need to test a hypothesis about the population variance or standard deviation. Assuming that the data are normally distributed, you use the χ^2 test for the variance or standard deviation to test whether the population variance or standard deviation is equal to a specified value. The [Section 12.7 online topic](#) discusses this test and illustrates its use.

12.8 Wilcoxon Signed Ranks Test: A Nonparametric Test for Two Related Populations

In Section 10.2, you used the paired t test to compare the means of two populations when you had repeated measures or matched samples. The paired t test assumes that the data are measured on an interval or a ratio scale and are normally distributed. If you cannot make these assumptions, you can use the nonparametric **Wilcoxon signed ranks test** to test for the median difference. The [Section 12.8 online topic](#) discusses this test and illustrates its use.

12.9 Friedman Rank Test: A Nonparametric Test for the Randomized Block Design

When analyzing a randomized block design, sometimes the data consists only of ranks within each block. Other times, you cannot assume that the data from each of the c groups are from normally distributed populations. In these situations, you can use the **Friedman rank test**. The [Section 12.9 online topic](#) discusses this test and illustrates its use.

USING STATISTICS

Avoiding Guesswork About Resort Guests, Revisited

In the Using Statistics scenario, you were the manager of T.C. Resort Properties, a collection of five upscale hotels located on two tropical islands. To assess the quality of services being provided by your hotels, guests are encouraged to complete a satisfaction survey when they check out or via email after they check out. You analyzed the data from these surveys to determine the overall satisfaction with the services provided, the likelihood that the guests will return to the hotel, and the reasons given by some guests for not wanting to return.

On one island, T.C. Resort Properties operates the Beachcomber and Windsurfer hotels. You performed a chi-square test for the difference in two proportions and concluded that a greater proportion of guests are willing to return to the Beachcomber Hotel than to the Windsurfer. On the other island, T.C. Resort Properties operates the Golden Palm, Palm Royale, and Palm Princess hotels. To see if guest satisfaction was the same among the three hotels, you



Matuso1812/Shutterstock

performed a chi-square test for the differences among more than two proportions. The test confirmed that the three proportions are not equal, and guests seem to be most likely to return to the Palm Royale and least likely to return to the Golden Palm.

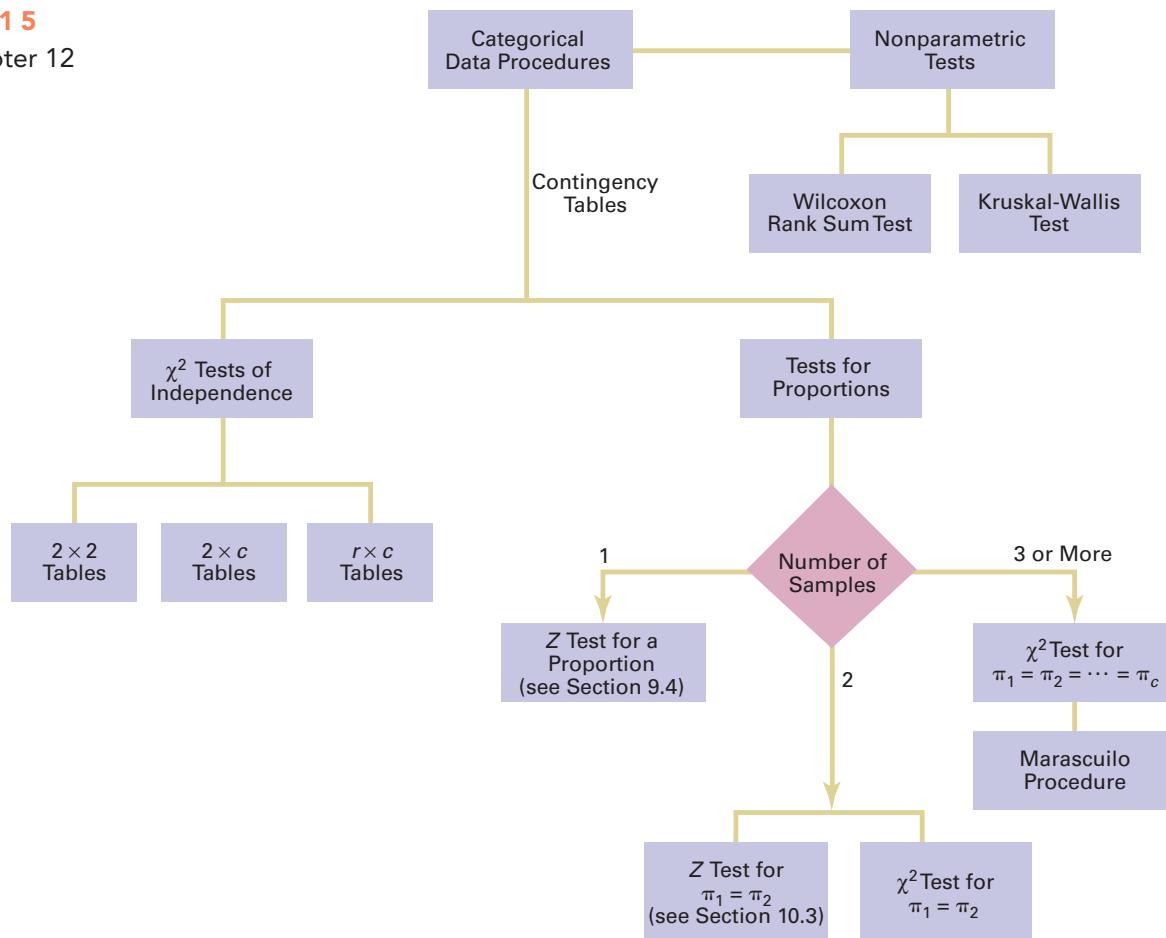
In addition, you investigated whether the reasons given for not returning to the Golden Palm, Palm Royale, and Palm Princess were unique to a certain hotel or common to all three hotels. By performing a chi-square test of independence, you determined that the reasons given for wanting to return or not depended on the hotel where the guests had been staying. By examining the observed and expected frequencies, you concluded that guests were more satisfied with the price at the Golden Palm and were much more satisfied with the location of the Palm Princess. Guest satisfaction with room accommodations was not significantly different among the three hotels.

SUMMARY

Figure 12.15 presents a roadmap for this chapter. First, you used hypothesis testing for analyzing categorical data from two independent samples and from more than two independent samples. In addition, the rules of probability from Section 4.2 were extended to the hypothesis of independence in the joint responses to two categorical variables.

You also studied two nonparametric tests. You used the Wilcoxon rank sum test when the assumptions of the t test for two independent samples were violated and the Kruskal-Wallis test when the assumptions of the one-way ANOVA F test were violated.

FIGURE 12.15
Roadmap of Chapter 12



REFERENCES

- Conover, W. J. *Practical Nonparametric Statistics*, 3rd ed. New York: Wiley, 2000.
- Daniel, W. W. *Applied Nonparametric Statistics*, 2nd ed. Boston: PWS Kent, 1990.
- Dixon, W. J., and F. J. Massey, Jr. *Introduction to Statistical Analysis*, 4th ed. New York: McGraw-Hill, 1983.
- Hollander, M., and D. A. Wolfe. *Nonparametric Statistical Methods*, 2nd ed. New York: Wiley, 1999.
- Lewontin, R. C., and J. Felsenstein. "Robustness of Homogeneity Tests in $2 \times n$ Tables," *Biometrics*, 21(March 1965): 19–33.
- Marascuilo, L. A. "Large-Sample Multiple Comparisons," *Psychological Bulletin*, 65(1966): 280–290.
- Marascuilo, L. A., and M. McSweeney. *Nonparametric and Distribution-Free Methods for the Social Sciences*. Monterey, CA: Brooks/Cole, 1977.
- Microsoft Excel 2013*. Redmond, WA: Microsoft Corp., 2012.
- Minitab Release 16*. State College, PA: Minitab Inc., 2010.
- Winer, B. J., D. R. Brown, and K. M. Michels. *Statistical Principles in Experimental Design*, 3rd ed. New York: McGraw-Hill, 1989.

KEY EQUATIONS

χ^2 Test for the Difference Between Two Proportions

$$\chi_{STAT}^2 = \sum_{all\ cells} \frac{(f_o - f_e)^2}{f_e} \quad (12.1)$$

Computing the Estimated Overall Proportion for Two Groups

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{X}{n} \quad (12.2)$$

Computing the Estimated Overall Proportion for c Groups

$$\bar{p} = \frac{X_1 + X_2 + \dots + X_c}{n_1 + n_2 + \dots + n_c} = \frac{X}{n} \quad (12.3)$$

Critical Range for the Marascuilo Procedure

$$\text{Critical range} = \sqrt{\chi_{\alpha}^2} \sqrt{\frac{p_j(1-p_j)}{n_j} + \frac{p_{j'}(1-p_{j'})}{n_{j'}}} \quad (12.4)$$

Computing the Expected Frequency

$$f_e = \frac{\text{Row total} \times \text{Column total}}{n} \quad (12.5)$$

Checking the Rankings

$$T_1 + T_2 = \frac{n(n+1)}{2} \quad (12.6)$$

Large-Sample Wilcoxon Rank Sum Test

$$Z_{STAT} = \frac{T_1 - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_2(n+1)}{12}}} \quad (12.7)$$

Kruskal-Wallis Rank Test for Differences Among c Medians

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^c \frac{T_j^2}{n_j} \right] - 3(n+1) \quad (12.8)$$

KEY TERMS

chi-square (χ^2) distribution 450

chi-square (χ^2) test for the difference between two proportions 449

chi-square (χ^2) test of independence 462

expected frequency (f_e) 449

Kruskal-Wallis rank test 473

Marascuilo procedure 459

nonparametric methods 467

observed frequency (f_o) 449

$2 \times c$ contingency table 455

2×2 contingency table 448

two-way contingency table 448

Wilcoxon rank sum test 467

CHECKING YOUR UNDERSTANDING

12.51 Under what conditions should you use the χ^2 test to determine whether there is a difference between the proportions of two independent populations?

12.52 Under what conditions should you use the χ^2 test to determine whether there is a difference among the proportions of more than two independent populations?

12.53 Under what conditions should you use the χ^2 test of independence?

12.54 Under what conditions should you use the Wilcoxon rank sum test instead of the t test for the difference between the means?

12.55 Under what conditions should you use the Kruskal-Wallis rank test instead of the one-way ANOVA?

CHAPTER REVIEW PROBLEMS

12.56 Undergraduate students at Miami University in Oxford, Ohio, were surveyed in order to evaluate the effect of gender and price on purchasing a pizza from Pizza Hut. Students were told to suppose that they were planning to have a large two-topping pizza

delivered to their residence that evening. The students had to decide between ordering from Pizza Hut at a reduced price of \$8.49 (the regular price for a large two-topping pizza from the Oxford Pizza Hut at the time was \$11.49) and ordering a pizza from a

different pizzeria. The results from this question are summarized in the following contingency table:

PIZZERIA			
GENDER	Pizza Hut	Other	Total
Female	4	13	17
Male	6	12	18
Total	10	25	35

- a. Using a 0.05 level of significance, is there evidence of a difference between males and females in their pizzeria selection?
- b. What is your answer to (a) if nine of the male students selected Pizza Hut and nine selected another pizzeria?

A subsequent survey evaluated purchase decisions at other prices. These results are summarized in the following contingency table:

PRICE				
PIZZERIA	\$8.49	\$11.49	\$14.49	Total
Pizza Hut	10	5	2	17
Other	25	23	27	75
Total	35	28	29	92

- c. Using a 0.05 level of significance and using the data in the second contingency table, is there evidence of a difference in pizzeria selection based on price?
- d. Determine the p -value in (c) and interpret its meaning.

12.57 What social media tools do marketers commonly use? The “2012 Social Media Marketing Industry Report” by Social Media Examiner (socialmediaexaminer.com) surveyed the percentage of marketers who commonly use an indicated social media tool. Surveyed were both B2B marketers, marketers that focus primarily on attracting businesses, and B2C marketers, marketers that primarily target consumers. Suppose the survey was based on 500 B2B marketers and 500 B2C marketers and yielded the results in the following table. (Data extracted from bit.ly/QmMxPa.)

BUSINESS FOCUS		
SOCIAL MEDIA TOOL	B2B	B2C
Facebook	87%	96%
Twitter	84%	80%
LinkedIn	87%	59%
YouTube or other video	56%	59%

For each social media tool, at the 0.05 level of significance, determine whether there is a difference between B2B marketers and B2C marketers in the proportion who used each social media tool.

12.58 A company is considering an organizational change involving the use of self-managed work teams. To assess the attitudes of employees of the company toward this change, a sample

of 400 employees is selected and asked whether they favor the institution of self-managed work teams in the organization. Three responses are permitted: favor, neutral, or oppose. The results of the survey, cross-classified by type of job and attitude toward self-managed work teams, are summarized as follows:

TYPE OF JOB	SELF-MANAGED WORK TEAMS			Total
	Favor	Neutral	Oppose	
Hourly worker	108	46	71	225
Supervisor	18	12	30	60
Middle management	35	14	26	75
Upper management	24	7	9	40
Total	185	79	136	400

- a. At the 0.05 level of significance, is there evidence of a relationship between attitude toward self-managed work teams and type of job?

The survey also asked respondents about their attitudes toward instituting a policy whereby an employee could take one additional vacation day per month without pay. The results, cross-classified by type of job, are as follows:

TYPE OF JOB	VACATION TIME WITHOUT PAY			Total
	Favor	Neutral	Oppose	
Hourly worker	135	23	67	225
Supervisor	39	7	14	60
Middle management	47	6	22	75
Upper management	26	6	8	40
Total	247	42	111	400

- b. At the 0.05 level of significance, is there evidence of a relationship between attitude toward vacation time without pay and type of job?

12.59 A company that produces and markets continuing education programs on DVDs for the educational testing industry has traditionally mailed advertising to prospective customers. A market research study was undertaken to compare two approaches: mailing a sample DVD upon request that contained highlights of the full DVD and sending an email containing a link to a website from which sample material could be downloaded. Of those who responded to either the mailing or the email, the results were as follows in terms of purchase of the complete DVD:

PURCHASED	TYPE OF MEDIA USED		
	Mailing	Email	Total
Yes	26	11	37
No	227	247	474
Total	253	258	511

- At the 0.05 level of significance, is there evidence of a difference in the proportion of DVDs purchased on the basis of the type of media used?
- On the basis of the results of (a), which type of media should the company use in the future? Explain the rationale for your decision.

The company also wanted to determine which of three sales approaches should be used to generate sales among those who either requested the sample DVD by mail or downloaded the sample DVD but did not purchase the full DVD: (1) targeted email, (2) a DVD that contained additional features, or (3) a telephone call to prospective customers. The 474 respondents who did not initially purchase the full DVD were randomly assigned to one of the three sales approaches. The results, in terms of purchases of the full-program DVD, are as follows:

ACTION	SALES APPROACH				Total
	Targeted Email	More Complete DVD	Telephone Call		
Purchase	5	17	18		40
Don't purchase	153	141	140		434
Total	158	158	158		474

- At the 0.05 level of significance, is there evidence of a difference in the proportion of DVDs purchased on the basis of the sales strategy used?
- On the basis of the results of (c), which sales approach do you think the company should use in the future? Explain the rationale for your decision.

CASES FOR CHAPTER 12

Managing Ashland MultiComm Services

PHASE 1

Reviewing the results of its research, the marketing department team concluded that a segment of Ashland households might be interested in a discounted trial subscription to the AMS *3-For-All* cable/phone/Internet service. The team decided to test various discounts before determining the type of discount to offer during the trial period. It decided to conduct an experiment using three types of discounts plus a plan that offered no discount during the trial period:

- No discount for the *3-For-All* cable/phone/Internet service. Subscribers would pay \$24.99 per week for the *3-For-All* cable/phone/Internet service during the 90-day trial period.
- Moderate discount for the *3-For-All* cable/phone/Internet service. Subscribers would pay \$19.99 per week for the *3-For-All* cable/phone/Internet service during the 90-day trial period.
- Substantial discount for the *3-For-All* cable/phone/Internet service. Subscribers would pay \$14.99 per week for the *3-For-All* cable/phone/Internet service during the 90-day trial period.
- Discount restaurant card. Subscribers would be given a special card providing a discount of 15% at selected restaurants in Ashland during the trial period.

Each participant in the experiment was randomly assigned to a discount plan. A random sample of 100 subscribers to each plan during the trial period was tracked to determine how many would continue to subscribe to the *3-For-All* service after the trial period. Table AMS12.1 summarizes the results.

TABLE AMS12.1

Number of Subscribers Who Continue Subscriptions After Trial Period with Four Discount Plans

CONTINUE SUBSCRIPTIONS AFTER TRIAL PERIOD	DISCOUNT PLANS				Total
	No Discount	Moderate Discount	Substantial Discount	Restaurant Card	
Yes	24	30	38	51	143
No	76	70	62	49	257
Total	100	100	100	100	400

- Analyze the results of the experiment. Write a report to the team that includes your recommendation for which discount plan to use. Be prepared to discuss the limitations and assumptions of the experiment.

PHASE 2

The marketing department team discussed the results of the survey presented in Chapter 8, on pages 301–302. The team realized that the evaluation of individual questions was providing only limited information. In order to further understand the market for the *3-For-All* cable/phone/Internet service, the data were organized in the following contingency tables:

HAS AMS TELEPHONE SERVICE	HAS AMS INTERNET SERVICE		
	Yes	No	Total
Yes	55	28	83
No	207	128	335
Total	262	156	418

		DISCOUNT TRIAL		
TYPE OF SERVICE		Yes	No	Total
Basic		8	156	164
Enhanced		32	222	254
Total		40	378	418

		METHOD FOR CURRENT SUBSCRIPTION					
DISCOUNT		Toll-Free Phone	AMS Website	Direct Mail Reply Card	Good Tunes & More	Other	Total
Yes		11	21	5	1	2	40
No		219	85	41	9	24	378
Total		230	106	46	10	26	418

WATCHES PREMIUM OR ON-DEMAND SERVICES					
TYPE OF SERVICE	Several Times a Week		Almost Never		Total
	Almost Every Day	Times a Week	Almost Never	Never	
Basic	2	5	127	30	164
Enhanced	12	30	186	26	254
Total	14	35	313	56	418

WATCHES PREMIUM OR ON-DEMAND SERVICES					
DISCOUNT	Several Times a Week		Almost Never		Total
	Almost Every Day	Times a Week	Almost Never	Never	
Yes	4	5	27	4	40
No	10	30	286	52	378
Total	14	35	313	56	418

Digital Case

Apply your knowledge of testing for the difference between two proportions in this Digital Case, which extends the T.C. Resort Properties Using Statistics scenario of this chapter.

As T.C. Resort Properties seeks to improve its customer service, the company faces new competition from SunLow Resorts. SunLow has recently opened resort hotels on the islands where T.C. Resort Properties has its five hotels. SunLow is currently advertising that a random survey of 300 customers revealed that about 60% of the customers preferred its “Concierge Class” travel reward program over the T.C. Resorts “TCRewards Plus” program.

Open and review **ConciergeClass.pdf**, an electronic brochure that describes the Concierge Class program and

METHOD FOR CURRENT SUBSCRIPTION							
GOLD CARD	Toll-Free Phone		AMS Website	Direct Mail Reply Card	Good Tunes & More	Other	Total
	Yes	No	Total	Yes	No	Total	
Yes	10	220	230	20	86	106	38
No	220	10	230	41	9	10	380
Total	230	14	418	46	10	26	418

2. Analyze the results of the contingency tables. Write a report for the marketing department team, discussing the marketing implications of the results for Ashland MultiComm Services.

compares it to the T.C. Resorts program. Then answer the following questions:

- Are the claims made by SunLow valid?
- What analyses of the survey data would lead to a more favorable impression about T.C. Resort Properties?
- Perform one of the analyses identified in your answer to step 2.
- Review the data about the T.C. Resort Properties customers presented in this chapter. Are there any other questions that you might include in a future survey of travel reward programs? Explain.

Sure Value Convenience Stores

You work in the corporate office for a nationwide convenience store franchise that operates nearly 10,000 stores. The per-store daily customer count (i.e., the mean number of customers in a store in one day) has been steady,

at 900, for some time. To increase the customer count, the chain is considering cutting prices for coffee beverages. Management needs to determine how much prices can be cut in order to increase the daily customer count without

reducing the gross margin on coffee sales too much. You decide to carry out an experiment in a sample of 24 stores where customer counts have been running almost exactly at the national average of 900. In 6 of the stores, a small coffee will be \$0.59, in another 6 stores the price will be \$0.69, in a third group of 6 stores, the price will be \$0.79, and in a fourth group of 6 stores, the price will now be \$0.89. After

four weeks, the daily customer count in the stores is stored in **CoffeeSales**.

At the 0.05 level of significance, is there evidence of a difference in the median daily customer count based on the price of a small coffee? What price should the stores sell the coffee for?

CardioGood Fitness

Return to the CardioGood Fitness case first presented on page 83. The data for this case are stored in **CardioGood Fitness**.

1. Determine whether differences exist in the median age in years, education in years, annual household income (\$), number of times the customer plans to use the treadmill each week, and the number of miles the customer expects to walk or run each week based on the product purchased (TM195, TM498, TM798).
2. Determine whether differences exist in the relationship status (single or partnered), and the self-rated fitness based on the product purchased (TM195, TM498, TM798).
3. Write a report to be presented to the management of CardioGood Fitness, detailing your findings.

More Descriptive Choices Follow-Up

Follow up the “Using Statistics: More Descriptive Choices, Revisited” on page 138 by using the data that are stored in **Retirement Funds** to:

1. Determine whether there is a difference between the growth and value funds in the median three-year return percentages, five-year return percentages, and ten-year return percentages.
2. Determine whether there is a difference between the small, mid-cap, and large market cap funds in the median three-year return percentages, five-year return percentages, and ten-year return percentages.
3. Determine whether there is a difference in risk based on market cap, a difference in rating based on market cap, a difference in risk based on type of fund, and a difference in rating based on type of fund.
4. Write a report summarizing your findings.

Clear Mountain State Student Surveys

1. The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. It creates and distributes a survey of 14 questions and receives responses from 62 undergraduates, which it stores in **UndergradSurvey**.
 - a. Construct contingency tables using gender, major, plans to go to graduate school, and employment status. (You need to construct six tables, taking two variables at a time.) Analyze the data at the 0.05 level of significance to determine whether any significant relationships exist among these variables.
 - b. At the 0.05 level of significance, is there evidence of a difference between males and females in median grade point average, expected starting salary, number of social networking sites registered for, age, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?
 - c. At the 0.05 level of significance, is there evidence of a difference between students who plan to go to graduate school and those who do not plan to go to graduate school in median grade point average, expected starting salary, number of social networking sites registered for, age, spending on textbooks and supplies,

- text messages sent in a week, and the wealth needed to feel rich?
- d. At the 0.05 level of significance, is there evidence of a difference based on academic major, in median expected starting salary, number of social networking sites registered for, age, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?
- e. At the 0.05 level of significance, is there evidence of a difference based on graduate school intention in median grade point average, expected starting salary, number of social networking sites registered for, age, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?
2. The dean of students at CMSU has learned about the undergraduate survey and has decided to undertake a similar survey for graduate students at CMSU. She creates and distributes a survey of 14 questions and receives responses from 44 graduate students, which she stores them in **GradSurvey**. For these data, at the 0.05 level of significance:
- a. Construct contingency tables using gender, undergraduate major, graduate major, and employment status. (You need to construct six tables, taking two variables at a time.) Analyze the data to determine whether any significant relationships exist among these variables.
- b. Is there evidence of a difference between males and females in the median age, undergraduate grade point average, graduate grade point average, expected salary upon graduation, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?
- c. Is there evidence of a difference based on undergraduate major in the median age, undergraduate grade point average, graduate grade point average, expected salary upon graduation, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?
- d. Is there evidence of a difference based on graduate major in the median age, undergraduate grade point average, graduate grade point average, expected salary upon graduation, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?
- e. Is there evidence of a difference based on employment status in the median age, undergraduate grade point average, graduate grade point average, expected salary upon graduation, spending on textbooks and supplies, text messages sent in a week, and the wealth needed to feel rich?

CHAPTER 12 EXCEL GUIDE

EG12.1 CHI-SQUARE TEST for the DIFFERENCE BETWEEN TWO PROPORTIONS

Key Technique Use the **CHISQ.INV.RT**(*level of significance, degrees of freedom*) function to compute the critical value and use the **CHISQ.DIST.RT**(*chi-square test statistic, degrees of freedom*) function to compute the *p*-value.

Example Perform this chi-square test for the two-hotel guest satisfaction data shown in Figure 12.3 on page 452.

PHStat Use Chi-Square Test for Differences in Two Proportions.

For the example, select **PHStat → Two-Sample Tests (Summarized Data) → Chi-Square Test for Differences in Two Proportions**. In the procedure's dialog box, enter **0.05** as the **Level of Significance**, enter a **Title**, and click **OK**. In the new worksheet:

1. Read the yellow note about entering values and then press the **Delete** key to delete the note.
2. Enter **Hotel** in cell **B4** and **Choose Again?** in cell **A5**.
3. Enter **Beachcomber** in cell **B5** and **Windsurfer** in cell **C5**.
4. Enter **Yes** in cell **A6** and **No** in cell **A7**.
5. Enter **163, 64, 154**, and **108** in cells **B6, B7, C6**, and **C7**, respectively.

In-Depth Excel Use the COMPUTE worksheet of the Chi-Square workbook as a template.

The worksheet already contains the Table 12.2 two-hotel guest satisfaction data. For other problems, change the **Observed Frequencies** cell counts and row and column labels in rows 4 through 7.

Read the **SHORT TAKES** for Chapter 12 for an explanation of the formulas found in the COMPUTE worksheet (shown in the **COMPUTE_FORMULAS worksheet**). If you are using an older Excel version, use the **COMPUTE_Older** worksheet.

EG12.2 CHI-SQUARE TEST for DIFFERENCES AMONG MORE THAN TWO PROPORTIONS

Key Technique Use the **CHISQ.INV.RT** and **CHISQ.DIST.RT** functions to compute the critical value and the *p*-value, respectively.

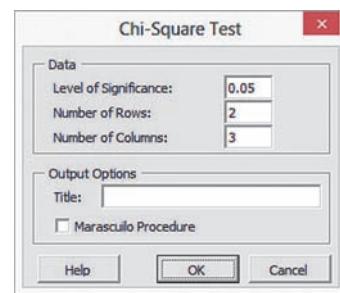
Example Perform this chi-square test for the three-hotel guest satisfaction data shown in Figure 12.6 on page 458.

PHStat Use Chi-Square Test.

For the example, select **PHStat → Multiple-Sample Tests → Chi-Square Test**. In the procedure's dialog box (shown in right column):

1. Enter **0.05** as the **Level of Significance**.

2. Enter **2** as the **Number of Rows**.
3. Enter **3** as the **Number of Columns**.
4. Enter a **Title** and click **OK**.



In the new worksheet:

5. Read the yellow note instructions about entering values and then press the **Delete** key to delete the note.
6. Enter the Table 12.6 data (see page 456), including row and column labels, in rows 4 through 7. The **#DIV/0!** error messages will disappear when you finish entering all the table data.

In-Depth Excel Use the **ChiSquare2x3 worksheet** of the **Chi-Square Worksheets** workbook as a model.

The worksheet already contains the Table 12.6 guest satisfaction data (see page 456). For other 2×3 problems, change the **Observed Frequencies** cell counts and row and column labels in rows 4 through 7. For 2×4 problems, use the **ChiSquare2x4 worksheet** and change the **Observed Frequencies** cell counts and row and column labels in that worksheet. For 2×5 problems, use the **ChiSquare2x5 worksheet** and change the **Observed Frequencies** cell counts and row and column labels in that worksheet.

The formulas that are found in the **ChiSquare2x3** workbook (shown in the **ChiSquare2x3_FORMULAS worksheet**) are similar to the formulas found in the COMPUTE worksheet of the Chi-Square workbook (see the previous section). If you use an Excel version older than Excel 2010, use the **ChiSquare2x3_Older** worksheet.

The Marascuilo Procedure

Key Technique Use formulas to compute the absolute differences and the critical range.

Example Perform the **Marascuilo Procedure** for the guest satisfaction survey that is shown in Figure 12.7 on page 460.

PHStat Modify the **PHStat** instructions of the previous section. In step 4, check **Marascuilo Procedure** in addition to entering a **Title** and clicking **OK**.

In-Depth Excel Use the **Marascuilo2x3** of the **Chi-Square Worksheets** workbook as a template.

The worksheet requires no entries or changes to use. For 2×4 problems, use the **Marascuilo2x4 worksheet** and for 2×5 problems, use the **Marascuilo2x5 worksheet**.

Read the SHORT TAKES for Chapter 12 for an explanation of the formulas found in the Marascuilo2x3 worksheet (shown in the **Marascuilo2x3_FORMULAS worksheet**). If you use an Excel version older than Excel 2010, use the **Marascuilo2x3_OLDER worksheet**.

EG12.3 CHI-SQUARE TEST of INDEPENDENCE

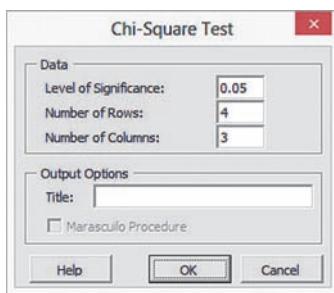
Key Technique Use the **CHISQ.INV.RT** and **CHISQ.DIST.RT** functions to compute the critical value and the *p*-value, respectively.

Example Perform this chi-square test for the primary reason for not returning to hotel data that is shown in Figure 12.10 on page 465.

PHStat Use Chi-Square Test.

For the example, select **PHStat** → **Multiple-Sample Tests** → **Chi-Square Test**. In the procedure's dialog box (shown below):

1. Enter **0.05** as the **Level of Significance**.
2. Enter **4** as the **Number of Rows**.
3. Enter **3** as the **Number of Columns**.
4. Enter a **Title** and click **OK**.



In the new worksheet:

5. Read the yellow note about entering values and then press the **Delete** key to delete the note.
6. Enter the Table 12.9 data on page 462, including row and column labels, in rows 4 through 9. The **#DIV/0!** error messages will disappear when you finish entering all of the table data.

In-Depth Excel Use the **ChiSquare4x3 worksheet** of the **Chi-Square Worksheets** workbook as a model.

The worksheet already contains the Table 12.9 primary reason for not returning to hotel data (see page 462). For other 4×3 problems, change the **Observed Frequencies** cell counts and row and column labels in rows 4 through 9. For 3×4 problems, use the **ChiSquare3x4 worksheet**. For 4×3 problems, use the **ChiSquare4x3 worksheet**. For 7×3 problems, use

the **ChiSquare7x3 worksheet**. For 8×3 problems, use the **ChiSquare8x3 worksheet**. For each of these other worksheets, enter the contingency table data for the problem in the Observed Frequencies area.

Read the SHORT TAKES for Chapter 12 to the Calculations area in columns G through I (not shown in Figure 12.10). The formulas found in the COMPUTE worksheet (shown in the **COMPUTE_FORMULAS worksheet**) are similar to those in the other chi-square worksheets discussed in this Excel Guide.

If you use an Excel version older than Excel 2010, use the **ChiSquare4x3_OLDER worksheet**.

EG12.4 WILCOXON RANK SUM TEST: a NONPARAMETRIC METHOD for TWO INDEPENDENT POPULATIONS

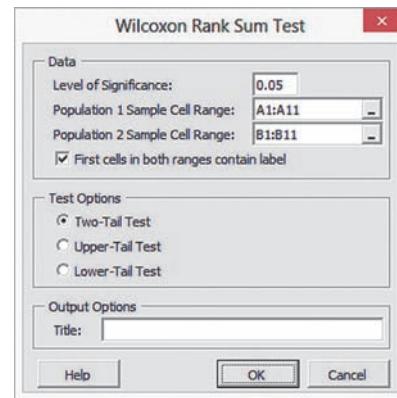
Key Technique Use the **NORM.S.INV(*level of significance*)** function to compute the upper and lower critical values and use **NORM.S.DIST(*absolute value of the Z test statistic*)** as part of a formula to compute the *p*-value. For unsummarized data, use the **COUNTIF** and **SUMIF** functions (see Appendix Section F.4) to compute the sample sizes and the sum of ranks for a sample, respectively.

Example Perform the Figure 12.12 Wilcoxon rank sum test for the cola sales for the two different end-cap locations.

PHStat Use Wilcoxon Rank Sum Test.

For the example, open to the **DATA worksheet** of the **Cola** workbook. Select **PHStat** → **Two-Sample Tests (Unsummarized Data)** → **Wilcoxon Rank Sum Test**. In the procedure's dialog box (shown below):

1. Enter **0.05** as the **Level of Significance**.
2. Enter **A1:A11** as the **Population 1 Sample Cell Range**.
3. Enter **B1:B11** as the **Population 2 Sample Cell Range**.
4. Check **First cells in both ranges contain label**.
5. Click **Two-Tail Test**.
6. Enter a **Title** and click **OK**.



The procedure creates a **SortedRanks** worksheet that contains the sorted ranks in addition to the worksheet shown in Figure 12.12. Both of these worksheets are discussed in the following **In-Depth Excel** instructions.

In-Depth Excel Use the **COMPUTE worksheet** of the **Wilcoxon workbook** as a template.

The worksheet already contains data and formulas to use the unsummarized data for the example. For other problems that use unsummarized data, first open to the **SortedRanks worksheet** and enter the sorted values for both groups in stacked format. Use column A for the sample names and column B for the sorted values. Assign a rank for each value and enter the ranks in column C of the same worksheet. Then open to the COMPUTE worksheet (or the similar COMPUTE_ALL worksheet, if performing a one-tail test) and edit the formulas in cells B7, B8, B10, and B11.

For problems with summarized data, overwrite the formulas that compute the **Sample Size** and **Sum of Ranks** in the cell range **B7:B11**, with the values for these statistics.

Open to the **COMPUTE_ALL_FORMULAS worksheet** to view all formulas in the COMPUTE_ALL worksheet. If you use an Excel version older than Excel 2010, use the **COMPUTE_ALL_OLDER** worksheet for all tests.

EG12.5 KRUSKAL-WALLIS RANK TEST: a NONPARAMETRIC METHOD for the ONE-WAY ANOVA

Key Technique Use the **CHISQ.INV.RT(*level of significance, number of groups - 1*)** function to compute the critical value and use the **CHISQ.DIST.RT(*H test statistic, number of groups - 1*)** function to compute the *p*-value. For unsummarized data, use the **COUNTIF** and **SUMIF** functions (see Appendix Section F.4) to compute the sample sizes and the sum of ranks for a sample, respectively.

Example Perform the Figure 12.14 Kruskal-Wallis rank test for differences among the four median mobile electronics sales for four in-store locations that is shown on page 475.

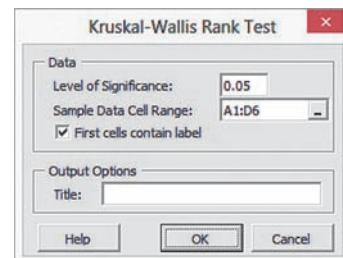
PHStat Use **Kruskal-Wallis Rank Test**.

For the example, open to the **DATA worksheet** of the **Second Experiment workbook**. Select **PHStat → Multiple-Sample Tests → Kruskal-Wallis Rank Test**. In the procedure's dialog box (shown in right column):

1. Enter **0.05** as the **Level of Significance**.
2. Enter **A1:D6** as the **Sample Data Cell Range**.

3. Check **First cells contain label**.

4. Enter a **Title** and click **OK**.



The procedure creates a **SortedRanks worksheet** that contains sorted ranks in addition to the worksheet shown in Figure 12.14 on page 475. Both of these worksheets are discussed in the following **In-Depth Excel** instructions.

In-Depth Excel Use the **KruskalWallis4 worksheet** of the **Kruskal-Wallis Worksheets workbook** as a template.

The worksheet already contains the data and formulas to use the unsummarized data for the example. For other problems with four groups and unsummarized data, first open to the **SortedRanks worksheet** and enter the sorted values for both groups in stacked format. Use column A for the sample names and column B for the sorted values. Assign ranks for each value and enter the ranks in column C of the same worksheet. Also paste your unsummarized stacked data in columns, starting with column E. (The row 1 cells, starting with cell E1, are used to identify each group.) Then open to the KruskalWallis4 worksheet and edit the formulas in columns E and F.

For other problems with four groups and summarized data, open to the **KruskalWallis4 worksheet** and overwrite the formulas that display the group names and compute the **Sample Size** and **Sum of Ranks** in columns D, E, and F with the values for these statistics. For other problems with three groups, use the similar **KruskalWallis3 worksheet**.

Open to the **KruskalWallis4_FORMULAS worksheet** to view all formulas in the **KruskalWallis4 worksheet**. If you use an Excel version older than Excel 2010, use the **KruskalWallis4_OLDER** or **KruskalWallis3_OLDER** worksheets.

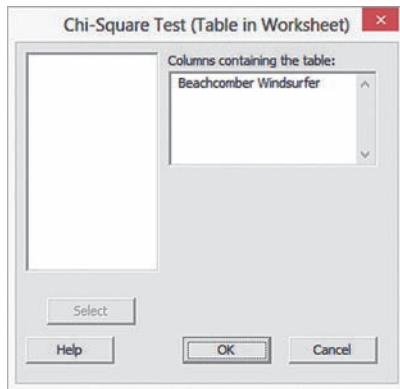
CHAPTER 12 MINITAB GUIDE

MG12.1 CHI-SQUARE TEST for the DIFFERENCE BETWEEN TWO PROPORTIONS

Use **Chi-Square Test (Two-Way Table in Worksheet)** (requires summarized data).

For example, to perform the Figure 12.3 test for the two-hotel guest satisfaction data on page 452, open to the **Two-Hotel Survey worksheet**. Select **Stat → Tables → Chi-Square Test (Two-Way Table in Worksheet)**. In the Chi-Square Test (Table in Worksheet) dialog box (shown below):

1. Double-click **C2 Beachcomber** in the variables list to add **Beachcomber** to the **Columns containing the table** box.
2. Double-click **C3 Windsurfer** in the variables list to add **Windsurfer** to the **Columns containing the table** box.
3. Click **OK**.



Minitab can also perform a chi-square test for the difference between two proportions using unsummarized data. Use the Section MG2.1 instructions for using **Cross Tabulation and Chi-Square** to create contingency tables (see page 96), replacing step 4 with these steps 4 through 7:

4. Click **Chi-Square**.

In the Cross Tabulation - Chi-Square dialog box:

5. Select **Chi-Square analysis, Expected cell counts, and Each cell's contribution to the Chi-Square statistic**.
6. Click **OK**.
7. Back in the original dialog box, click **OK**.

MG12.2 CHI-SQUARE TEST for DIFFERENCES AMONG MORE THAN TWO PROPORTIONS

Use **Chi-Square Test (Two-Way Table in Worksheet)** (requires summarized data).

Use modified Section MG2.1 instructions on the page 96 for using **Cross Tabulation and Chi-Square** to perform the chi-square test with unsummarized data. See Section MG12.1 for detailed instructions.

To perform the Figure 12.6 test for the three-hotel guest satisfaction data on page 458, open to the **Three-Hotel Survey worksheet**, select **Stat → Tables → Chi-Square Test (Two-Way Table in Worksheet)**. In the Chi-Square Test (Table in Worksheet) dialog box, enter the names of columns 2 through 4 in the **Columns containing the table** box and click **OK**.

The Marascuilo Procedure

There are no Minitab Guide instructions for this procedure.

MG12.3 CHI-SQUARE TEST of INDEPENDENCE

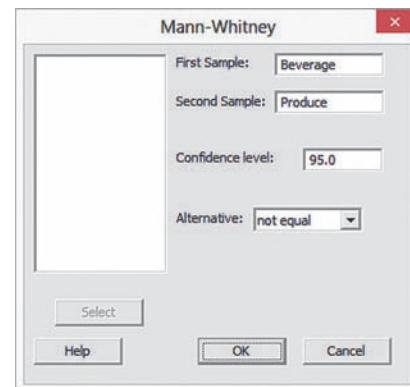
Use the Section MG2.1 instructions for either **Chi-Square Test (Two-Way Table in Worksheet)** for summarized data or the (modified) **Cross Tabulation and Chi-Square** for unsummarized data to perform this test.

MG12.4 WILCOXON RANK SUM TEST: a NONPARAMETRIC METHOD for TWO INDEPENDENT POPULATIONS

Use **Mann-Whitney** to perform a test equivalent to the Wilcoxon rank sum test.

For example, to perform the Figure 12.12 test for the cola sales for the two different end-cap locations on page 470, open to the **Cola worksheet**. Select **Stat → Nonparametrics → Mann-Whitney**. In the Mann-Whitney dialog box (shown below):

1. Double-click **C1 Beverage** in the variables list to add **Beverage** in the **First Sample** box.
2. Double-click **C2 Produce** in the variables list to add **Produce** in the **Second Sample** box.
3. Enter **95.0** in the **Confidence level** box.
4. Select **not equal** in the **Alternative** drop-down list.
5. Click **OK**.



MG12.5 KRUSKAL-WALLIS RANK TEST: a NONPARAMETRIC METHOD for the ONE-WAY ANOVA

Use **Kruskal-Wallis**.

For example, to perform the Figure 12.14 Kruskal-Wallis rank test for differences among the four median mobile electronics sales for four in-store locations on page 474, open to the **Second Experiment Stacked worksheet**. Select **Stat → Nonparametrics → Kruskal-Wallis**. In the Kruskal-Wallis dialog box (shown at right):

1. Double-click **C2 Sales** in the variables list to add **Sales** in the **Response** box.
2. Double-click **C1 Location** in the variables list to add **Location** in the **Factor** box.
3. Click **OK**.



CHAPTER 13

Simple Linear Regression

CONTENTS

- 13.1 Types of Regression Models
- 13.2 Determining the Simple Linear Regression Equation
- VISUAL EXPLORATIONS: Exploring Simple Linear Regression Coefficients**
- 13.3 Measures of Variation
- 13.4 Assumptions of Regression
- 13.5 Residual Analysis
- 13.6 Measuring Autocorrelation: The Durbin-Watson Statistic
- 13.7 Inferences About the Slope and Correlation Coefficient
- 13.8 Estimation of Mean Values and Prediction of Individual Values
- 13.9 Potential Pitfalls in Regression

Six Steps for Avoiding the Potential Pitfalls

USING STATISTICS: Knowing Customers at Sunflowers Apparel, Revisited

CHAPTER 13 EXCEL GUIDE

CHAPTER 13 MINITAB GUIDE

OBJECTIVES

- To learn to use regression analysis to predict the value of a dependent variable based on the value of an independent variable
- Understand the meaning of the regression coefficients b_0 and b_1
- To learn to evaluate the assumptions of regression analysis and what to do if the assumptions are violated
- To make inferences about the slope and correlation coefficient
- To estimate mean values and predict individual values

USING STATISTICS

Knowing Customers at Sunflowers Apparel

Having survived recent economic slowdowns that have diminished their competitors, Sunflowers Apparel, a chain of upscale fashion stores for women, is in the midst of a companywide review that includes researching the factors that make their stores successful. Until recently, Sunflowers managers did not use data analysis to help select where to open stores, relying instead on subjective factors, such as the availability of an inexpensive lease or the perception that a particular location seemed ideal for one of their stores.

As the new director of planning, you have already consulted with marketing data firms that specialize in identifying and classifying groups of consumers. Based on such preliminary analyses, you have already tentatively discovered that the profile of Sunflowers shoppers may not only be the upper middle class long suspected of being the chain's clientele but may also include younger, aspirational families with young children, and, surprisingly, urban hipsters that set trends and are mostly single.

You seek to develop a systematic approach that will lead to making better decisions during the site-selection process. As a starting point, you have asked one marketing data firm to collect and organize data for the number of people in the identified groups of interest who live within a fixed radius of each store. You believe that the greater numbers of profiled customers contribute to store sales, and you want to explore the possible use of this relationship in the decision-making process. How can you use statistics so that you can forecast the annual sales of a proposed store based on the number of profiled customers that reside within a fixed radius of a Sunflowers store?



Fotolia

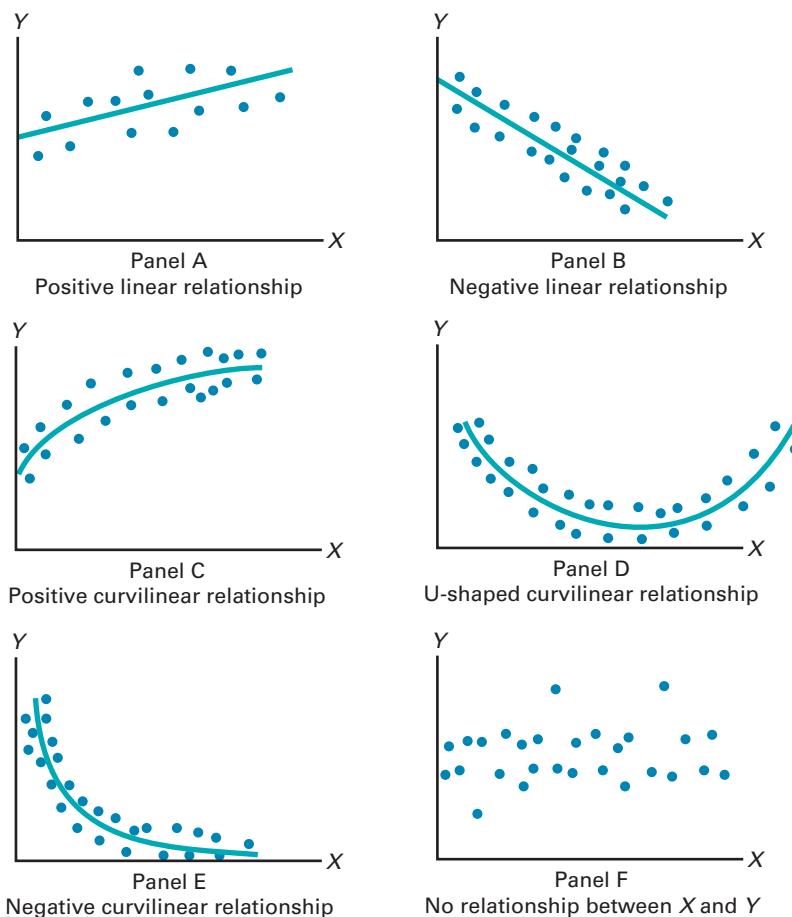
In this chapter and the next two chapters, you learn **regression analysis** techniques that help uncover relationships between variables. Regression analysis leads to selection of a **model** that expresses how one or more **independent variables** can be used to predict the value of another variable, called the **dependent variable**. Regression models identify the type of mathematical relationship that exists between a dependent variable and an independent variable, thereby enabling you to quantify the effect that a change in the independent variable has on the dependent variable. Models also help you identify unusual values that may be outliers (see page 578).

This chapter discusses **simple linear regression** models that use a single numerical independent variable, X , to predict the numerical dependent variable, Y . (Chapters 14 and 15 discuss *multiple* regression models that use several independent variables to predict the dependent variable.) In the Sunflowers scenario, your initial belief reflects a possible simple linear regression model in which the number of profiled customers would be the single numerical independent variable, X , being used to predict the annual sales of the store, the dependent variable, Y .

13.1 Types of Regression Models

Using a scatter plot (also known as scatter diagram) to visualize the X and Y variables, a technique introduced in Section 2.5 on page 65, can help suggest a starting point for regression analysis. The scatter plots in Figure 13.1 illustrates six possible relationships between an X and Y variable.

FIGURE 13.1
Six types of relationships found in scatter plots



In Panel A, values of Y are generally increasing linearly as X increases. This panel is similar to Figure 13.3 on page 494, which illustrates the positive relationship between the number of profiled customers of the store and the store's annual sales for the Sunflowers Apparel women's clothing store chain.

Panel B is an example of a negative linear relationship. As X increases, the values of Y are generally decreasing. An example of this type of relationship might be the price of a particular product and the amount of sales. As the price charged for the product increases, the amount of sales may tend to decrease.

Panel C shows a positive curvilinear relationship between X and Y . The values of Y increase as X increases, but this increase tapers off beyond certain values of X . An example of a positive curvilinear relationship might be the age and maintenance cost of a machine. As a machine gets older, the maintenance cost may rise rapidly at first but then level off beyond a certain number of years.

Panel D shows a U-shaped relationship between X and Y . As X increases, at first Y generally decreases; but as X continues to increase, Y not only stops decreasing but actually increases above its minimum value. An example of this type of relationship might be entrepreneurial activity and levels of economic development as measured by GDP per capita. Entrepreneurial activity occurs more in the least and most developed countries.

Panel E illustrates an exponential relationship between X and Y . In this case, Y decreases very rapidly as X first increases, but then it decreases much less rapidly as X increases further. An example of an exponential relationship could be the value of an automobile and its age. The value drops drastically from its original price in the first year, but it decreases much less rapidly in subsequent years.

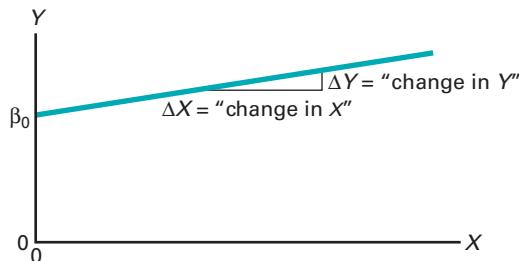
Finally, Panel F shows a set of data in which there is very little or no relationship between X and Y . High and low values of Y appear at each value of X .

Simple Linear Regression Models

Although scatter plots provide preliminary analysis, more sophisticated statistical procedures determine the most appropriate model for a set of variables. Simple linear regression models represent the simplest relationship of a straight-line or **linear relationship**. Figure 13.2 illustrates this relationship.

FIGURE 13.2

A straight-line relationship



Equation (13.1) expresses this relationship mathematically by defining the simple linear regression model.

SIMPLE LINEAR REGRESSION MODEL

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (13.1)$$

where

β_0 = Y intercept for the population

β_1 = slope for the population

ε_i = random error in Y for observation i

Y_i = dependent variable (sometimes referred to as the **response variable**) for observation i

X_i = independent variable (sometimes referred to as the predictor, or **explanatory variable**) for observation i

The $Y_i = \beta_0 + \beta_1 X_i$ portion of the simple linear regression model expressed in Equation (13.1) is a straight line. The **slope** of the line, β_1 , represents the expected change in Y per unit change in X . It represents the mean amount that Y changes (either positively or negatively) for a one-unit change in X . The **Y intercept**, β_0 , represents the mean value of Y when X equals 0. The last component of the model, ε_i , represents the random error in Y for each observation, i . In other words, ε_i is the vertical distance of the actual value of Y_i above or below the expected value of Y_i on the line.

13.2 Determining the Simple Linear Regression Equation

In the Sunflowers Apparel scenario on page 491, the business objective of the director of planning is to forecast annual sales for all new stores, based on the number of profiled customers who live no more than 30 minutes from a Sunflowers store. To examine the relationship between the number of profiled customers (in millions) who live within a fixed radius from a Sunflowers store and its annual sales (\$millions), data were collected from a sample of 14 stores. Table 13.1 shows the organized data, which are stored in [SiteSelection](#).

TABLE 13.1

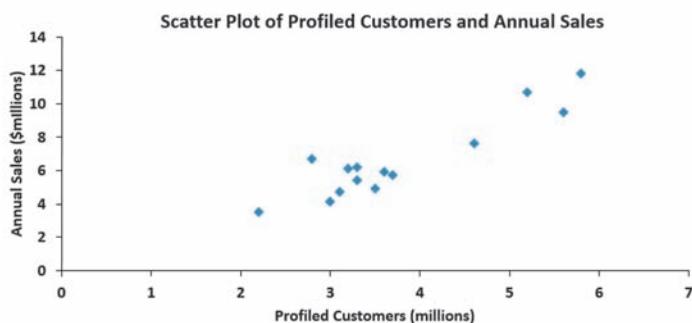
Number of Profiled Customers (in millions) and Annual Sales (in \$millions) for a Sample of 14 Sunflowers Apparel Stores

Store	Profiled Customers (millions)	Annual Sales (\$millions)	Store	Profiled Customers (millions)	Annual Sales (\$millions)
1	3.7	5.7	8	3.1	4.7
2	3.6	5.9	9	3.2	6.1
3	2.8	6.7	10	3.5	4.9
4	5.6	9.5	11	5.2	10.7
5	3.3	5.4	12	4.6	7.6
6	2.2	3.5	13	5.8	11.8
7	3.3	6.2	14	3.0	4.1

Figure 13.3 displays the scatter plot for the data in Table 13.1. Observe the increasing relationship between profiled customers (X) and annual sales (Y). As the number of profiled customers increases, annual sales increase approximately as a straight line (superimposed on scatter plot). Thus, you can assume that a straight line provides a useful mathematical model of this relationship. Now you need to determine the specific straight line that is the *best fit* to these data.

FIGURE 13.3

Scatter plot for the Sunflowers Apparel data



The Least-Squares Method

In the preceding section, a statistical model is hypothesized to represent the relationship between two variables—number of profiled customers and sales—in the entire population of Sunflowers Apparel stores. However, as shown in Table 13.1, the data are collected from a random sample of stores. If certain assumptions are valid (see Section 13.4), you can use the sample Y intercept, b_0 , and the sample slope, b_1 , as estimates of the respective population parameters, β_0 and β_1 . Equation (13.2) uses these estimates to form the **simple linear regression equation**. This straight line is often referred to as the **prediction line**.



► Student Tip

In mathematics, the symbol b is often used for the Y intercept instead of b_0 and the symbol m is often used for the slope instead of b_1 .

SIMPLE LINEAR REGRESSION EQUATION: THE PREDICTION LINE

The predicted value of Y equals the Y intercept plus the slope multiplied by the value of X .

$$\hat{Y}_i = b_0 + b_1 X_i \quad (13.2)$$

where

\hat{Y}_i = predicted value of Y for observation i

X_i = value of X for observation i

b_0 = sample Y intercept

b_1 = sample slope

Equation (13.2) requires you to determine two **regression coefficients**— b_0 (the sample Y intercept) and b_1 (the sample slope). The most common approach to finding b_0 and b_1 is using the least-squares method. This method minimizes the sum of the squared differences between the actual values (Y_i) and the predicted values (\hat{Y}_i), using the simple linear regression equation [i.e., the prediction line; see Equation (13.2)]. This sum of squared differences is equal to



► Student Tip

Although the solutions to Examples 13.3 and 13.4 (on pages 498–499 and 504–505, respectively) present the formulas for computing these values (and others), you should always consider using software to compute the values of the terms discussed in this chapter.

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Because $\hat{Y}_i = b_0 + b_1 X_i$,

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n [Y_i - (b_0 + b_1 X_i)]^2$$

Because this equation has two unknowns, b_0 and b_1 , the sum of squared differences depends on the sample Y intercept, b_0 , and the sample slope, b_1 . The **least-squares method** determines the values of b_0 and b_1 that minimize the sum of squared differences around the prediction line. Any values for b_0 and b_1 other than those determined by the least-squares method result in a greater sum of squared differences between the actual values (Y_i) and the predicted values (\hat{Y}_i).

Figure 13.4 presents the worksheet for the simple linear regression model for the Table 13.1 Sunflowers Apparel data. In this figure, Excel labels b_0 as Intercept and Minitab labels b_0 as Constant. They each label b_1 as Profiled Customers.

FIGURE 13.4

Excel and Minitab simple linear regression models for the Sunflowers Apparel data

A	B	C	D	E	F	G
Simple Linear Regression						
Regression Statistics						
Multiple R	0.9208					
R Square	0.8479					
Adjusted R Square	0.8352					
Standard Error	0.9993					
Observations	14					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	66.7854	66.7854	66.8792	0.0000	
Residual	12	11.9832	0.9986			
Total	13	78.7686				
Coefficients						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-1.2088	0.9949	-1.2151	0.2477	-3.3765	0.9588
Profiled Customers	2.0742	0.2536	8.1780	0.0000	1.5216	2.6268
Unusual Observations						
	Profiled	Annual				
Obs	Customers	Sales	Fit	SE Fit	Residual	St Resid
3	2.80	6.700	4.599	0.365	2.101	2.26R
R denotes an observation with a large standardized residual.						
Predicted Values for New Observations						
New Obs	Fit	SE Fit	95% CI		95% PI	
1	7.088	0.273	(6.493, 7.682)		(4.831, 9.345)	
Values of Predictors for New Observations						
	Profiled					
New Obs	Customers					
1	4.00					

In Figure 13.4, observe that $b_0 = -1.2088$ and $b_1 = 2.0742$. Using Equation (13.2) on page 495, the prediction line for these data is

$$\hat{Y}_i = -1.2088 + 2.0742X_i$$

Student Tip

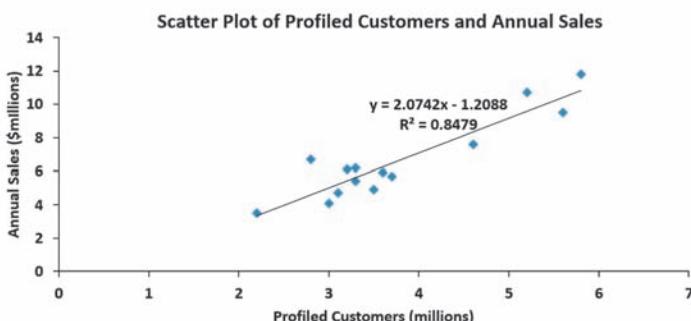
Remember that a positive slope means that as X increases, Y is predicted to increase. A negative slope means that as X increases, Y is predicted to decrease.

The slope, b_1 , is $+2.0742$. This means that for each increase of 1 unit in X , the predicted mean value of Y is estimated to increase by 2.0742 units. In other words, for each increase of 1.0 million profiled customers within 30 minutes of the store, the predicted mean annual sales are estimated to increase by \$2.0742 million. Thus, the slope represents the portion of the annual sales that are estimated to vary according to the number of profiled customers.

The Y intercept, b_0 , is -1.2088 . The Y intercept represents the predicted value of Y when X equals 0. Because the number of profiled customers of the store cannot be 0, this Y intercept has little or no practical interpretation. Also, the Y intercept for this example is outside the range of the observed values of the X variable, and therefore interpretations of the value of b_0 should be made cautiously. Figure 13.5 displays the actual values and the prediction line.

FIGURE 13.5

Scatter plot and prediction line for Sunflowers Apparel data



Example 13.1 illustrates a situation in which there is a direct interpretation for the Y intercept, b_0 .

EXAMPLE 13.1

Interpreting the Y Intercept, b_0 , and the Slope, b_1

A statistics professor wants to use the number of hours a student studies for a statistics final exam (X) to predict the final exam score (Y). A regression model is fit based on data collected from a class during the previous semester, with the following results:

$$\hat{Y}_i = 35.0 + 3X_i$$

What is the interpretation of the Y intercept, b_0 , and the slope, b_1 ?

SOLUTION The Y intercept $b_0 = 35.0$ indicates that when the student does not study for the final exam, the predicted mean final exam score is 35.0. The slope $b_1 = 3$ indicates that for each increase of one hour in studying time, the predicted change in the mean final exam score is $+3.0$. In other words, the final exam score is predicted to increase by a mean of 3 points for each one-hour increase in studying time.

Return to the Sunflowers Apparel scenario on page 491. Example 13.2 illustrates how you use the prediction line to predict the annual sales.

EXAMPLE 13.2**Predicting Annual Sales Based on Number of Profiled Customers**

Use the prediction line to predict the annual sales for a store with 4 million profiled customers.

SOLUTION You can determine the predicted value of annual sales by substituting $X = 4$ (millions of profiled customers) into the simple linear regression equation:

$$\hat{Y}_i = -1.2088 + 2.0742X_i$$

$$\hat{Y}_i = -1.2088 + 2.0742(4) = 7.0879 \text{ or } \$7,087,900$$

Thus, a store with 4 million profiled customers has predicted mean annual sales of \$7,087,900.

Predictions in Regression Analysis: Interpolation Versus Extrapolation

When using a regression model for prediction purposes, you should consider only the **relevant range** of the independent variable in making predictions. This relevant range includes all values from the smallest to the largest X used in developing the regression model. Hence, when predicting Y for a given value of X , you can interpolate within this relevant range of the X values, but you should not extrapolate beyond the range of X values. When you use the number of profiled customers to predict annual sales, the number of profiled customers (in millions) varies from 2.2 to 5.8 (see Table 13.1 on page 494). Therefore, you should predict annual sales *only* for stores that have between 2.2 and 5.8 million profiled customers. Any prediction of annual sales for stores outside this range assumes that the observed relationship between sales and the number of profiled customers for stores that have between 2.2 and 5.8 million profiled customers is the same as for stores outside this range. For example, you cannot extrapolate the linear relationship beyond 5.8 million profiled customers in Example 13.2. It would be improper to use the prediction line to forecast the sales for a new store that has 8 million profiled customers because the relationship between sales and the number of profiled customers may have a point of diminishing returns. If that is true, as the number of profiled customers increases beyond 5.8 million, the effect on sales may become smaller and smaller.

Computing the Y Intercept, b_0 , and the Slope, b_1

For small data sets, you can use a hand calculator to compute the least-squares regression coefficients. Equations (13.3) and (13.4) give the values of b_0 and b_1 , which minimize

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n [Y_i - (b_0 + b_1 X_i)]^2$$

COMPUTATIONAL FORMULA FOR THE SLOPE, b_1

$$b_1 = \frac{SSXY}{SSX} \quad (13.3)$$

where

$$SSXY = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}$$

$$SSX = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}$$

(continued)

COMPUTATIONAL FORMULA FOR THE Y INTERCEPT, b_0

$$b_0 = \bar{Y} - b_1 \bar{X} \quad (13.4)$$

where

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

EXAMPLE 13.3Computing the Y Intercept, b_0 , and the Slope, b_1

Compute the Y intercept, b_0 , and the slope, b_1 , for the Sunflowers Apparel data.

SOLUTION In Equations (13.3) and (13.4), five quantities need to be computed to determine b_1 and b_0 . These are n , the sample size; $\sum_{i=1}^n X_i$, the sum of the X values; $\sum_{i=1}^n Y_i$, the sum of the Y values; $\sum_{i=1}^n X_i^2$, the sum of the squared X values; and $\sum_{i=1}^n X_i Y_i$, the sum of the product of X and Y .

For the Sunflowers Apparel data, the number of profiled customers (X) is used to predict the annual sales (Y) in a store. Table 13.2 presents the computations of the sums needed for the site selection problem. The table also includes $\sum_{i=1}^n Y_i^2$, the sum of the squared Y values that will be used to compute SST in Section 13.3.

TABLE 13.2

Computations for the Sunflowers Apparel Data

Store	Profiled Customers (X)	Annual Sales (Y)	X^2	Y^2	XY
1	3.7	5.7	13.69	32.49	21.09
2	3.6	5.9	12.96	34.81	21.24
3	2.8	6.7	7.84	44.89	18.76
4	5.6	9.5	31.36	90.25	53.20
5	3.3	5.4	10.89	29.16	17.82
6	2.2	3.5	4.84	12.25	7.70
7	3.3	6.2	10.89	38.44	20.46
8	3.1	4.7	9.61	22.09	14.57
9	3.2	6.1	10.24	37.21	19.52
10	3.5	4.9	12.25	24.01	17.15
11	5.2	10.7	27.04	114.49	55.64
12	4.6	7.6	21.16	57.76	34.96
13	5.8	11.8	33.64	139.24	68.44
14	3.0	4.1	9.00	16.81	12.30
Totals	52.9	92.8	215.41	693.90	382.85


Student Tip

Coefficients computed manually with the assistance of handheld calculators may differ slightly because of rounding errors caused by the limited number of decimal places that your calculator might use.

Using Equations (13.3) and (13.4), you can compute b_0 and b_1 :

$$\begin{aligned} SSXY &= \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n} \\ &= 382.85 - \frac{(52.9)(92.8)}{14} \\ &= 382.85 - 350.65142 \\ &= 32.19858 \\ SSX &= \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n} \\ &= 215.41 - \frac{(52.9)^2}{14} \\ &= 215.41 - 199.88642 \\ &= 15.52358 \end{aligned}$$

With these values, compute b_1 :

$$\begin{aligned} b_1 &= \frac{SSXY}{SSX} \\ &= \frac{32.19858}{15.52358} \\ &= 2.07417 \end{aligned}$$

and:

$$\begin{aligned} \bar{Y} &= \frac{\sum_{i=1}^n Y_i}{n} = \frac{92.8}{14} = 6.62857 \\ \bar{X} &= \frac{\sum_{i=1}^n X_i}{n} = \frac{52.9}{14} = 3.77857 \end{aligned}$$

With these values, compute b_0 :

$$\begin{aligned} b_0 &= \bar{Y} - b_1 \bar{X} \\ &= 6.62857 - 2.07417(3.77857) \\ &= -1.2088265 \end{aligned}$$

VISUAL EXPLORATIONS**Exploring Simple Linear Regression Coefficients**

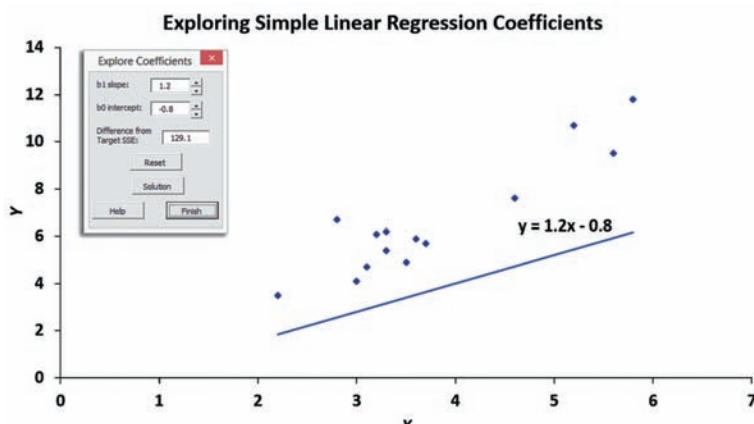
Open the **VE-Simple Linear Regression add-in workbook** to explore the coefficients. (See Appendix C to learn how you can download a copy of this workbook and Appendix Section D.5 before using this workbook.) When this workbook opens properly, it adds a **Simple Linear Regression** menu in either the Add-ins tab (Microsoft Windows) or the Apple menu bar (OS X).

To explore the effects of changing the simple linear regression coefficients, select **Simple Linear Regression → Explore Coefficients**. In the Explore Coefficients floating control panel (shown inset below), click the spinner buttons for **b_1 slope** (the slope of the prediction line) and **b_0 intercept** (the Y intercept of the prediction line) to change the prediction line. Using the visual feedback of the chart, try to create a prediction line that is as close as possible to the prediction line defined by the least-squares estimates. In other words, try to make the **Difference from Target SSE** value as small as possible. (See page 503 for an explanation of SSE.)

At any time, click **Reset** to reset the b_1 and b_0 values or **Solution** to reveal the prediction line defined by the least-squares method. Click **Finish** when you are finished with this exercise.

Using Your Own Regression Data

Select **Simple Linear Regression using your worksheet data** from the **Simple Linear Regression** menu to explore the simple linear regression coefficients using data you supply from a worksheet. In the procedure's dialog box, enter the cell range of your **YVariable Cell Range** and the cell range of your **XVariable Cell Range**. Click **First cells in both ranges contain a label**, enter a **Title**, and click **OK**. After the scatter plot appears onscreen, continue with the Explore Coefficients floating control panel as described in the left column.

**Problems for Section 13.2****LEARNING THE BASICS**

- 13.1** Fitting a straight line to a set of data yields the following prediction line:

$$\hat{Y}_i = 2 + 5X_i$$

- a. Interpret the meaning of the Y intercept, b_0 .
- b. Interpret the meaning of the slope, b_1 .
- c. Predict the value of Y for $X = 3$.

- 13.2** If the values of X in Problem 13.1 range from 2 to 25, should you use this model to predict the mean value of Y when X equals **a. 3?** **b. -3?** **c. 0?** **d. 24?**

- 13.3** Fitting a straight line to a set of data yields the following prediction line:

$$\hat{Y}_i = 16 - 0.5X_i$$

- a. Interpret the meaning of the Y intercept, b_0 .
- b. Interpret the meaning of the slope, b_1 .
- c. Predict the value of Y for $X = 6$.

APPLYING THE CONCEPTS

- SELF TEST** **13.4** The production of wine is a multibillion-dollar worldwide industry. In an attempt to develop a model of wine quality as judged by wine experts, data was collected from red wine variants of Portuguese “Vinho Verde” wine. (Data extracted from P. Cortez, Cerdeira, A., Almeida, F., Matos, T., and Reis, J., “Modeling Wine Preferences by Data Mining from Physiochemical Properties,” *Decision Support Systems*, 47, 2009, pp. 547–553 and bit.ly/9xKIEa.) A sample of 50 wines is stored in [VinhoVerde](#). Develop a simple linear regression model to predict wine quality, measured on a scale from 0 (very bad) to 10 (excellent), based on alcohol content (%).

- a. Construct a scatter plot.
For these data, $b_0 = -0.3529$ and $b_1 = 0.5624$.
- b. Interpret the meaning of the slope, b_1 , in this problem.
- c. Predict the mean wine quality for wines with a 10% alcohol content.
- d. What conclusion can you reach based on the results of (a)–(c)?

13.5 Zagat's publishes restaurant ratings for various locations in the United States. The file **Restaurants** contains the Zagat rating for food, décor, service, and the cost per person for a sample of 100 restaurants located in New York City and in a suburb of New York City. Develop a regression model to predict the cost per person, based on a variable that represents the sum of the ratings for food, décor, and service.

Sources: Extracted from *Zagat Survey 2013, New York City Restaurants*; and *Zagat Survey 2012–2013, Long Island Restaurants*.

- Construct a scatter plot.

For these data, $b_0 = -46.7718$ and $b_1 = 1.4963$.

- Assuming a linear relationship, use the least-squares method to compute the regression coefficients b_0 and b_1 .
- Interpret the meaning of the Y intercept, b_0 , and the slope, b_1 , in this problem.
- Predict the mean cost per person for a restaurant with a summated rating of 50.
- What should you tell the owner of a group of restaurants in this geographical area about the relationship between the summated rating and the cost of a meal?

13.6 The owner of a moving company typically has his most experienced manager predict the total number of labor hours that will be required to complete an upcoming move. This approach has proved useful in the past, but the owner has the business objective of developing a more accurate method of predicting labor hours. In a preliminary effort to provide a more accurate method, the owner has decided to use the number of cubic feet moved as the independent variable and has collected data for 36 moves in which the origin and destination were within the borough of Manhattan in New York City and in which the travel time was an insignificant portion of the hours worked. The data are stored in **Moving**.

- Construct a scatter plot.
- Assuming a linear relationship, use the least-squares method to determine the regression coefficients b_0 and b_1 .
- Interpret the meaning of the slope, b_1 , in this problem.
- Predict the mean labor hours for moving 500 cubic feet.
- What should you tell the owner of the moving company about the relationship between cubic feet moved and labor hours?

13.7 Starbucks Coffee Co. uses a data-based approach to improving the quality and customer satisfaction of its products. When survey data indicated that Starbucks needed to improve its package-sealing process, an experiment was conducted to determine the factors in the bag-sealing equipment that might be affecting the ease of opening the bag without tearing the inner liner of the bag. (Data extracted from L. Johnson and S. Burrows, "For Starbucks, It's in the Bag," *Quality Progress*, March 2011, pp. 17–23.) One factor that could affect the rating of the ability of the bag to resist tears was the plate gap on the bag-sealing equipment. Data were collected on 19 bags in which the plate gap was varied. The results are stored in **Starbucks**.

- Construct a scatter plot.
- Assuming a linear relationship, use the least-squares method to determine the regression coefficients b_0 and b_1 .
- Interpret the meaning of the slope, b_1 , in this problem.
- Predict the mean tear rating when the plate gap is equal to 0.
- What should you tell management of Starbucks about the relationship between the plate gap and the tear rating?

13.8 The value of a sports franchise is directly related to the amount of revenue that a franchise can generate. The file **BBRevenue2013** represents the value in 2013 (in \$millions) and the annual revenue (in \$millions) for the 30 Major League Baseball franchises. (Data extracted from www.forbes.com/mlb-valuations/list.) Suppose you want to develop a simple linear regression model to predict franchise value based on annual revenue generated.

- Construct a scatter plot.
- Use the least-squares method to determine the regression coefficients b_0 and b_1 .
- Interpret the meaning of b_0 and b_1 in this problem.
- Predict the mean value of a baseball franchise that generates \$250 million of annual revenue.
- What would you tell a group considering an investment in a major league baseball team about the relationship between revenue and the value of a team?

13.9 An agent for a residential real estate company in a suburb located outside of Washington, DC, has the business objective of developing more accurate estimates of the monthly rental cost for apartments. Toward that goal, the agent would like to use the size of an apartment, as defined by square footage to predict the monthly rental cost. The agent selects a sample of 48 one-bedroom apartments and collects and stores the data in **RentSilverSpring**.

- Construct a scatter plot.
- Use the least-squares method to determine the regression coefficients b_0 and b_1 .
- Interpret the meaning of b_0 and b_1 in this problem.
- Predict the mean monthly rent for an apartment that has 800 square feet.
- Why would it not be appropriate to use the model to predict the monthly rent for apartments that have 1,500 square feet?
- Your friends Jim and Jennifer are considering signing a lease for an one-bedroom apartment in this residential neighborhood. They are trying to decide between two apartments, one with 800 square feet for a monthly rent of \$1,130 and the other with 830 square feet for a monthly rent of \$1,410. Based on (a) through (d), which apartment do you think is a better deal?

13.10 A company that holds the DVD distribution rights to movies previously released only in theaters has the business objective of developing estimates of the sales revenue of DVDs. Toward this goal, a company analyst plans to use box office gross to predict DVD sales revenue. For 43 movies, the analyst collects the box office gross (in \$millions) in the year that they were released and the DVD revenue (in \$millions) in the following year and stores these data in **Movie**. (Data extracted from bit.ly/14uuPNB and bit.ly/HMh6Kx.)

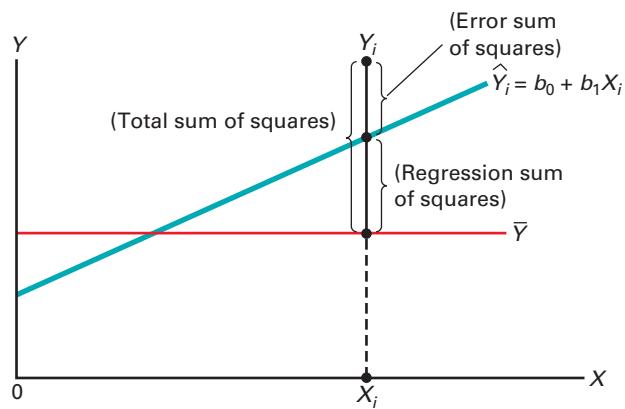
For these data,

- Construct a scatter plot.
- Assuming a linear relationship, use the least-squares method to determine the regression coefficients b_0 and b_1 .
- Interpret the meaning of the slope, b_1 , in this problem.
- Predict the mean sales revenue for a movie DVD that had a box office gross of \$100 million.
- What conclusions can you reach about predicting DVD revenue from movie gross?

13.3 Measures of Variation

When using the least-squares method to determine the regression coefficients for a set of data, you need to compute three measures of variation. The first measure, the **total sum of squares (SST)**, is a measure of variation of the Y_i values around their mean, \bar{Y} . The **total variation**, or total sum of squares, is subdivided into **explained variation** and **unexplained variation**. The explained variation, or **regression sum of squares (SSR)**, represents variation that is explained by the relationship between X and Y , and the unexplained variation, or **error sum of squares (SSE)**, represents variation due to factors other than the relationship between X and Y . Figure 13.6 shows the different measures of variation for a single Y_i value.

FIGURE 13.6
Measures of variation



Computing the Sum of Squares

The regression sum of squares (SSR) is based on the difference between \hat{Y}_i (the predicted value of Y from the prediction line) and \bar{Y} (the mean value of Y). The error sum of squares (SSE) represents the part of the variation in Y that is not explained by the regression. It is based on the difference between Y_i and \hat{Y}_i . The total sum of squares (SST) is equal to the regression sum of squares (SSR) plus the error sum of squares (SSE). Equations (13.5), (13.6), (13.7), and (13.8) define these measures of variation and the total sum of squares (SST).

MEASURES OF VARIATION IN REGRESSION

The total sum of squares (SST) is equal to the regression sum of squares (SSR) plus the error sum of squares (SSE).

$$SST = SSR + SSE \quad (13.5)$$

TOTAL SUM OF SQUARES (SST)

The total sum of squares (SST) is equal to the sum of the squared differences between each observed value of Y and the mean value of Y .

$$SST = \text{Total sum of squares}$$

$$= \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (13.6)$$

REGRESSION SUM OF SQUARES (SSR)

The regression sum of squares (*SSR*) is equal to the sum of the squared differences between each predicted value of *Y* and the mean value of *Y*.

$SSR = \text{Explained variation or regression sum of squares}$

$$= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad (13.7)$$

ERROR SUM OF SQUARES (SSE)

The error sum of squares (*SSE*) is equal to the sum of the squared differences between each observed value of *Y* and the predicted value of *Y*.

$SSE = \text{Unexplained variation or error sum of squares}$

$$= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (13.8)$$

Figure 13.7 shows the sum of squares portion of the Figure 13.4 results for the Sunflowers Apparel data. The total variation, *SST*, is equal to 78.7686. This amount is subdivided into the sum of squares explained by the regression (*SSR*), equal to 66.7854, and the sum of squares unexplained by the regression (*SSE*), equal to 11.9832. From Equation (13.5) on page 502:

$$SST = SSR + SSE$$

$$78.7686 = 66.7854 + 11.9832$$

FIGURE 13.7

Excel and Minitab sum of squares portion for the Sunflowers Apparel data

A	B	C	D	E	F	G	Analysis of Variance					
10	ANOVA	df	SS	MS	F	Significance F	Source	DF	SS	MS	F	P
11							Regression	1	66.785	66.785	66.88	0.000
12	Regression	1	66.7854	66.7854	66.8792	0.0000	Residual Error	12	11.983	0.999		
13	Residual	12	11.9832	0.9986			Total	13	78.769			
14	Total	13	78.7686									
15												
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%					
17	Intercept	-1.2088	0.9949	-1.2151	0.2477	-3.3765	0.9588					
18	Profiled Customers	2.0742	0.2536	8.1780	0.0000	1.5216	2.6265					

The Coefficient of Determination

By themselves, *SSR*, *SSE*, and *SST* provide little information. However, the ratio of the regression sum of squares (*SSR*) to the total sum of squares (*SST*) measures the proportion of variation in *Y* that is explained by the linear relationship of the independent variable *X* with the dependent variable *Y* in the regression model. This ratio, called the coefficient of determination, r^2 , is defined in Equation (13.9).

COEFFICIENT OF DETERMINATION

The coefficient of determination is equal to the regression sum of squares (i.e., explained variation) divided by the total sum of squares (i.e., total variation).

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST} \quad (13.9)$$

Student Tip

r^2 must be a value between 0 and 1. It cannot be negative.

The **coefficient of determination** measures the proportion of variation in Y that is explained by the variation in the independent variable X in the regression model.

For the Sunflowers Apparel data, with $SSR = 66.7854$, $SSE = 11.9832$, and $SST = 78.7686$,

$$r^2 = \frac{66.7854}{78.7686} = 0.8479$$

Therefore, 84.79% of the variation in annual sales is explained by the variability in the number of profiled customers. This large r^2 indicates a strong linear relationship between these two variables because the regression model has explained 84.79% of the variability in predicting annual sales. Only 15.21% of the sample variability in annual sales is due to factors other than what is accounted for by the linear regression model that uses the number of profiled customers.

Figure 13.8 presents the regression statistics table portion of the Figure 13.4 results for the Sunflowers Apparel data. This table contains the coefficient of determination.

FIGURE 13.8

Excel and Minitab regression statistics for the Sunflowers Apparel data

	A	B	Predictor	Coeff	SE Coef	T	P
Regression Statistics							
3			Constant	-1.2088	0.9949	-1.22	0.248
4	Multiple R	0.9208	Profiled Customers	2.0742	0.2536	8.18	0.000
5	R Square	0.8479					
6	Adjusted R Square	0.8352					
7	Standard Error	0.9993					
8	Observations	14					
<i>S = 0.999298 R-Sq = 84.8% R-Sq(adj) = 83.5%</i>							

EXAMPLE 13.4

Computing the Coefficient of Determination

Compute the coefficient of determination, r^2 , for the Sunflowers Apparel data.

SOLUTION You can compute SST , SSR , and SSE , which are defined in Equations (13.6), (13.7), and (13.8) on pages 502 and 503, by using Equations (13.10), (13.11), and (13.12).

COMPUTATIONAL FORMULA FOR SST

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} \quad (13.10)$$

COMPUTATIONAL FORMULA FOR SSR

$$\begin{aligned} SSR &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\ &= b_0 \sum_{i=1}^n Y_i + b_1 \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} \end{aligned} \quad (13.11)$$

COMPUTATIONAL FORMULA FOR SSE

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i Y_i \quad (13.12)$$

Using the summary results from Table 13.2 on page 498,

$$\begin{aligned} SST &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} \\ &= 693.9 - \frac{(92.8)^2}{14} \end{aligned}$$

$$\begin{aligned}
 &= 693.9 - 615.13142 \\
 &= 78.76858 \\
 SSR &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\
 &= b_0 \sum_{i=1}^n Y_i + b_1 \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n Y_i \right)^2}{n} \\
 &= (-1.2088265)(92.8) + (2.07417)(382.85) - \frac{(92.8)^2}{14} \\
 &= 66.7854 \\
 SSE &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\
 &= \sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i Y_i \\
 &= 693.9 - (-1.2088265)(92.8) - (2.07417)(382.85) \\
 &= 11.9832
 \end{aligned}$$

**Student Tip**

Coefficients computed manually with the assistance of handheld calculators may differ slightly.

Therefore,

$$r^2 = \frac{66.7854}{78.7686} = 0.8479$$

Standard Error of the Estimate

Although the least-squares method produces the line that fits the data with the minimum amount of prediction error, unless all the observed data points fall on a straight line, the prediction line is not a perfect predictor. Just as all data values cannot be expected to be exactly equal to their mean, neither can all the values in a regression analysis be expected to be located exactly on the prediction line. Figure 13.5 on page 496 illustrates the variability around the prediction line for the Sunflowers Apparel data. Notice that many of the observed values of Y fall near the prediction line, but none of the values are exactly on the line.

The **standard error of the estimate** measures the variability of the observed Y values from the predicted \hat{Y} values in the same way that the standard deviation in Chapter 3 measures the variability of each value around the sample mean. In other words, the standard error of the estimate is the standard deviation *around* the prediction line, whereas the standard deviation in Chapter 3 is the standard deviation *around* the sample mean. Equation (13.13) defines the standard error of the estimate, represented by the symbol S_{YX} .

STANDARD ERROR OF THE ESTIMATE

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} \quad (13.13)$$

where

Y_i = actual value of Y for a given X_i

\hat{Y}_i = predicted value of Y for a given X_i

SSE = error sum of squares

From Equation (13.8) and Figure 13.4 or Figure 13.7 on pages 495 or 503, $SSE = 11.9832$. Thus,

$$S_{YX} = \sqrt{\frac{11.9832}{14 - 2}} = 0.9993$$

This standard error of the estimate, equal to 0.9993 millions of dollars (i.e., \$999,300), is labeled Standard Error in the Figure 13.8 Excel results and S in the Minitab results. The standard error of the estimate represents a measure of the variation around the prediction line. It is measured in the same units as the dependent variable Y . The interpretation of the standard error of the estimate is similar to that of the standard deviation. Just as the standard deviation measures variability around the mean, the standard error of the estimate measures variability around the prediction line. For Sunflowers Apparel, the typical difference between actual annual sales at a store and the predicted annual sales using the regression equation is approximately \$999,300.

Problems for Section 13.3

LEARNING THE BASICS

13.11 How do you interpret a coefficient of determination, r^2 , equal to 0.80?

13.12 If $SSR = 36$ and $SSE = 4$, determine SST and then compute the coefficient of determination, r^2 , and interpret its meaning.

13.13 If $SSR = 66$ and $SST = 88$, compute the coefficient of determination, r^2 , and interpret its meaning.

13.14 If $SSE = 10$ and $SSR = 30$, compute the coefficient of determination, r^2 , and interpret its meaning.

13.15 If $SSR = 120$, why is it impossible for SST to equal 110?

APPLYING THE CONCEPTS

 **13.16** In Problem 13.4 on page 500, the percentage of alcohol was used to predict wine quality (stored in **VinhoVerde**). For those data, $SSR = 21.8677$ and $SST = 64.0000$.

- a. Determine the coefficient of determination, r^2 , and interpret its meaning.
- b. Determine the standard error of the estimate.
- c. How useful do you think this regression model is for predicting sales?

13.17 In Problem 13.5 on page 501, you used the summated rating to predict the cost of a restaurant meal (stored in **Restaurants**). For those data, $SSR = 9,740.0629$ and $SST = 17,844.75$.

- a. Determine the coefficient of determination, r^2 , and interpret its meaning.
- b. Determine the standard error of the estimate.
- c. How useful do you think this regression model is for predicting the cost of a restaurant meal?

13.18 In Problem 13.6 on page 501, an owner of a moving company wanted to predict labor hours, based on the cubic feet moved (stored in **Moving**). Using the results of that problem,

- a. determine the coefficient of determination, r^2 , and interpret its meaning.
- b. determine the standard error of the estimate.
- c. How useful do you think this regression model is for predicting labor hours?

13.19 In Problem 13.7 on page 501, you used the plate gap on the bag-sealing equipment to predict the tear rating of a bag of coffee (stored in **Starbucks**). Using the results of that problem,

- a. determine the coefficient of determination, r^2 , and interpret its meaning.
- b. determine the standard error of the estimate.
- c. How useful do you think this regression model is for predicting the tear rating based on the plate gap in the bag-sealing equipment?

13.20 In Problem 13.8 on page 501, you used annual revenues to predict the value of a baseball franchise (stored in **BBRevenue2013**). Using the results of that problem,

- a. determine the coefficient of determination, r^2 , and interpret its meaning.
- b. determine the standard error of the estimate.
- c. How useful do you think this regression model is for predicting the value of a baseball franchise?

13.21 In Problem 13.9 on page 501, an agent for a real estate company wanted to predict the monthly rent for one-bedroom apartments, based on the size of the apartment (stored in **Rent-SilverSpring**). Using the results of that problem,

- a. determine the coefficient of determination, r^2 , and interpret its meaning.
- b. determine the standard error of the estimate.
- c. How useful do you think this regression model is for predicting the monthly rent?
- d. Can you think of other variables that might explain the variation in monthly rent?

13.22 In Problem 13.10 on page 501, you used box office gross to predict DVD revenue (stored in **Movie**). Using the results of that problem,

- a. determine the coefficient of determination, r^2 , and interpret its meaning.
- b. determine the standard error of the estimate.
- c. How useful do you think this regression model is for predicting DVD revenue?
- d. Can you think of other variables that might explain the variation in DVD revenue?

13.4 Assumptions of Regression

When hypothesis testing and the analysis of variance were discussed in Chapters 9 through 12, the importance of the assumptions to the validity of any conclusions reached was emphasized. The assumptions necessary for regression are similar to those of the analysis of variance because both are part of the general category of *linear models* (reference 4).

The four **assumptions of regression** (known by the acronym LINE) are as follows:

- Linearity
- Independence of errors
- Normality of error
- Equal variance

The first assumption, **linearity**, states that the relationship between variables is linear. Relationships between variables that are not linear are discussed in Chapter 15.

The second assumption, **independence of errors**, requires that the errors (ε_i) be independent of one another. This assumption is particularly important when data are collected over a period of time. In such situations, the errors in a specific time period are sometimes correlated with those of the previous time period.

The third assumption, **normality**, requires that the errors (ε_i) be normally distributed at each value of X . Like the t test and the ANOVA F test, regression analysis is fairly robust against departures from the normality assumption. As long as the distribution of the errors at each level of X is not extremely different from a normal distribution, inferences about β_0 and β_1 are not seriously affected.

The fourth assumption, **equal variance**, or **homoscedasticity**, requires that the variance of the errors (ε_i) be constant for all values of X . In other words, the variability of Y values is the same when X is a low value as when X is a high value. The equal-variance assumption is important when making inferences about β_0 and β_1 . If there are serious departures from this assumption, you can use either data transformations or weighted least-squares methods (see reference 4).

13.5 Residual Analysis

Sections 13.2 and 13.3 developed a regression model using the least-squares method for the Sunflowers Apparel data. Is this the correct model for these data? Are the assumptions presented in Section 13.4 valid? **Residual analysis** visually evaluates these assumptions and helps you determine whether the regression model that has been selected is appropriate.

The **residual**, or estimated error value, e_i , is the difference between the observed (Y_i) and predicted (\hat{Y}_i) values of the dependent variable for a given value of X_i . A residual appears on a scatter plot as the vertical distance between an observed value of Y and the prediction line. Equation (13.14) defines the residual.

RESIDUAL

The residual is equal to the difference between the observed value of Y and the predicted value of Y .

$$e_i = Y_i - \hat{Y}_i \quad (13.14)$$

Evaluating the Assumptions

Recall from Section 13.4 that the four assumptions of regression (known by the acronym LINE) are linearity, independence, normality, and equal variance.

Student Tip

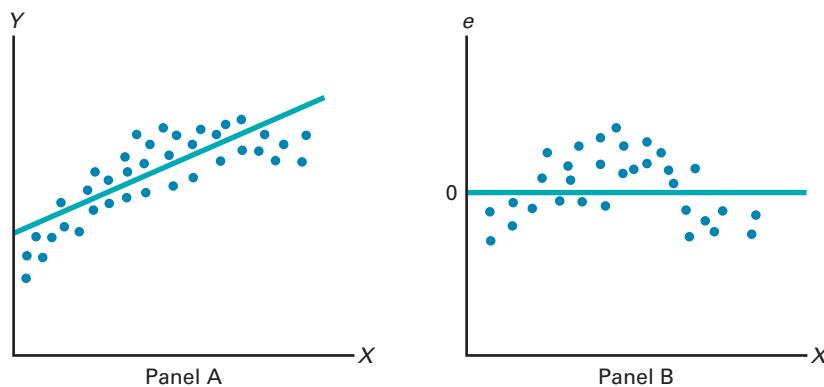
When there is no apparent pattern in the residual plot, the plot will look like a random scattering of points.

Linearity To evaluate linearity, you plot the residuals on the vertical axis against the corresponding X_i values of the independent variable on the horizontal axis. If the linear model is appropriate for the data, you will not see any apparent pattern in the plot. However, if the linear model is not appropriate, in the residual plot, there will be a relationship between the X_i values and the residuals, e_i .

You can see such a pattern in the residuals in Figure 13.9. Panel A shows a situation in which, although there is an increasing trend in Y as X increases, the relationship seems curvilinear because the upward trend decreases for increasing values of X . This effect is even more apparent in Panel B, where there is a clear relationship between X_i and e_i . By removing the linear trend of X with Y , the residual plot has exposed the lack of fit in the simple linear model more clearly than the scatter plot in Panel A. For these data, a quadratic or curvilinear model (see Section 15.1) is a better fit and should be used instead of the simple linear model.

FIGURE 13.9

Studying the appropriateness of the simple linear regression model



To determine whether the simple linear regression model for the Sunflowers Apparel data is appropriate, you need to determine the residuals. Figure 13.10 displays the predicted annual sales values and residuals for the Sunflowers Apparel data.

FIGURE 13.10

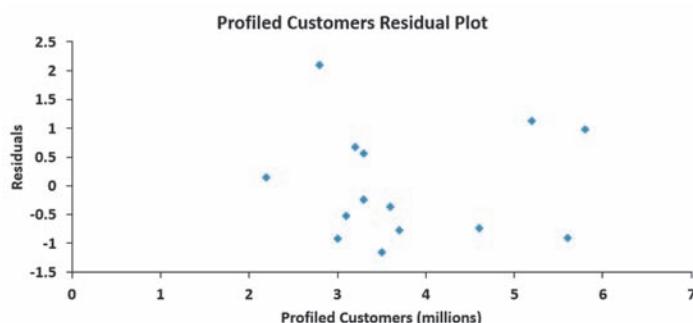
Table of residuals for the Sunflowers Apparel data

Store	Profiled Customers	Predicted Annual Sales	Annual Sales	Residuals
1	3.7	6.4656	5.7	-0.7656
2	3.6	6.2582	5.9	-0.3582
3	2.8	4.5988	6.7	2.1012
4	5.6	10.4065	9.5	-0.9065
5	3.3	5.6359	5.4	-0.2359
6	2.2	3.3543	3.5	0.1457
7	3.3	5.6359	6.2	0.5641
8	3.1	5.2211	4.7	-0.5211
9	3.2	5.4285	6.1	0.6715
10	3.5	6.0508	4.9	-1.1508
11	5.2	9.5769	10.7	1.1231
12	4.6	8.3324	7.6	-0.7324
13	5.8	10.8214	11.8	0.9786
14	3	5.0137	4.1	-0.9137

To assess linearity, you plot the residuals against the independent variable (number of profiled customers, in millions) in Figure 13.11. Although there is widespread scatter in the residual plot, there is no clear pattern or relationship between the residuals and X_i . The residuals appear to be evenly spread above and below 0 for different values of X . You can conclude that the linear model is appropriate for the Sunflowers Apparel data.

FIGURE 13.11

Plot of residuals against the profiled customers of a store for the Sunflowers Apparel data



Independence You can evaluate the assumption of independence of the errors by plotting the residuals in the order or sequence in which the data were collected. If the values of Y are part of a time series (see Section 2.5), a residual may sometimes be related to the residual that precedes it. If this relationship exists between consecutive residuals (which violates the assumption of independence), the plot of the residuals versus the time in which the data were collected will often show a cyclical pattern. Because the Sunflowers Apparel data were collected during the same time period, you do not need to evaluate the independence assumption for these data.

Normality You can evaluate the assumption of normality in the errors by constructing a histogram (see Section 2.4), using a stem-and-leaf display (see Section 2.4), a boxplot (see Section 3.3), or a normal probability plot (see Section 6.3). To evaluate the normality assumption for the Sunflowers Apparel data, Table 13.3 organizes the residuals into a frequency distribution and Figure 13.12 is a normal probability plot.

TABLE 13.3

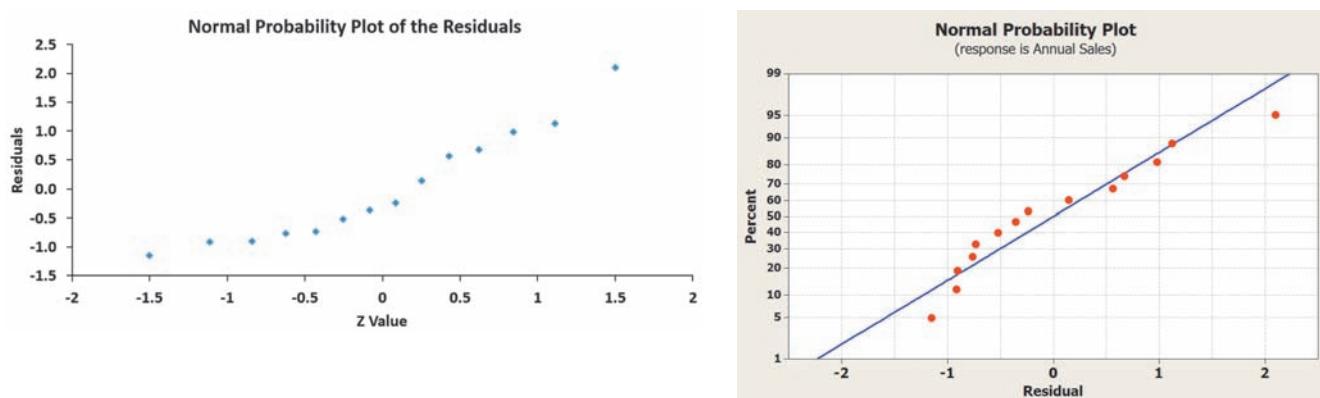
Frequency Distribution of 14 Residual Values for the Sunflowers Apparel Data

Residuals	Frequency
-1.25 but less than -0.75	4
-0.75 but less than -0.25	3
-0.25 but less than +0.25	2
+0.25 but less than +0.75	2
+0.75 but less than +1.25	2
+1.25 but less than +1.75	0
+1.75 but less than +2.25	1
	14

Although the small sample size makes it difficult to evaluate normality, from the normal probability plot of the residuals in Figure 13.12, the data do not appear to depart substantially from a normal distribution. The robustness of regression analysis with modest departures from normality enables you to conclude that you should not be overly concerned about departures from this normality assumption in the Sunflowers Apparel data.

FIGURE 13.12

Excel and Minitab normal probability plots of the residuals for the Sunflowers Apparel data



LEARN MORE

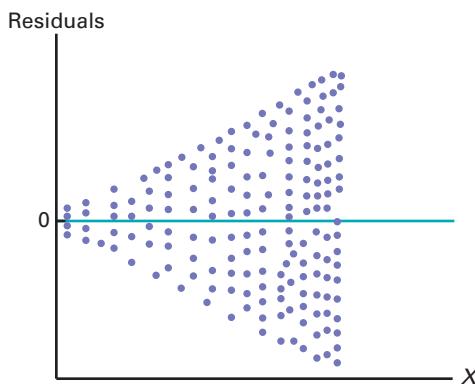
You can also test the assumption of equal variance by performing the White test (see reference 7). Learn more about this test in the [White Test online topic](#).

Equal Variance You can evaluate the assumption of equal variance from a plot of the residuals with X_i . You examine the plot to see if there is approximately the same amount of variation in the residuals at each value of X . For the Sunflowers Apparel data of Figure 13.11 on page 608, there do not appear to be major differences in the variability of the residuals for different X_i values. Thus, you can conclude that there is no apparent violation in the assumption of equal variance at each level of X .

To examine a case in which the equal-variance assumption is violated, observe Figure 13.13, which is a plot of the residuals with X_i for a hypothetical set of data. This plot is fan shaped because the variability of the residuals increases dramatically as X increases. Because this plot shows unequal variances of the residuals at different levels of X , the equal-variance assumption is invalid.

FIGURE 13.13

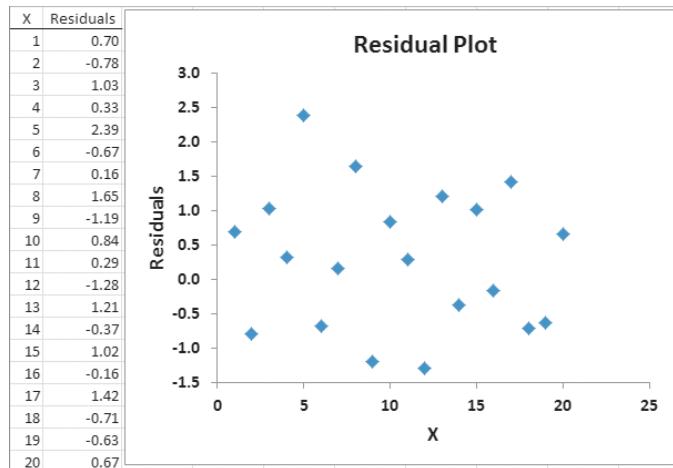
Violation of equal variance



Problems for Section 13.5

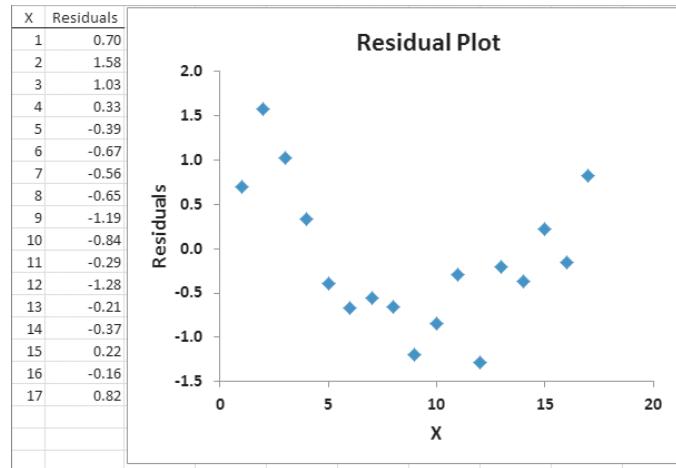
LEARNING THE BASICS

13.23 The following results provide the X values, residuals, and a residual plot from a regression analysis:



Is there any evidence of a pattern in the residuals? Explain.

13.24 The following results show the X values, residuals, and a residual plot from a regression analysis:



Is there any evidence of a pattern in the residuals? Explain.

APPLYING THE CONCEPTS

13.25 In Problem 13.5 on page 501, you used the summated rating to predict the cost of a restaurant meal. Perform a residual analysis for these data (stored in **Restaurants**). Evaluate whether the assumptions of regression have been seriously violated.

 **SELF Test** **13.26** In Problem 13.4 on page 500, you used the percentage of alcohol to predict wine quality. Perform a residual analysis for these data (stored in **VinhoVerde**). Evaluate whether the assumptions of regression have been seriously violated.

13.27 In Problem 13.7 on page 501, you used the plate gap on the bag-sealing equipment to predict the tear rating of a bag of coffee. Perform a residual analysis for these data (stored in **Starbucks**). Based on these results, evaluate whether the assumptions of regression have been seriously violated.

13.28 In Problem 13.6 on page 501, the owner of a moving company wanted to predict labor hours based on the cubic feet moved. Perform a residual analysis for these data (stored in **Moving**).

Based on these results, evaluate whether the assumptions of regression have been seriously violated.

13.29 In Problem 13.9 on page 501, an agent for a real estate company wanted to predict the monthly rent for one-bedroom apartments, based on the size of the apartments. Perform a residual analysis for these data (stored in **RentSilverSpring**). Based on these results, evaluate whether the assumptions of regression have been seriously violated.

13.30 In Problem 13.8 on page 501, you used annual revenues to predict the value of a baseball franchise. Perform a residual analysis for these data (stored in **BBRevenue2013**). Based on these results, evaluate whether the assumptions of regression have been seriously violated.

13.31 In Problem 13.10 on page 501, you used box office gross to predict DVD revenue. Perform a residual analysis for these data (stored in **Movie**). Based on these results, evaluate whether the assumptions of regression have been seriously violated.

13.6 Measuring Autocorrelation: The Durbin-Watson Statistic

One of the basic assumptions of the regression model is the independence of the errors. This assumption is sometimes violated when data are collected over sequential time periods because a residual at any one time period sometimes is similar to residuals at adjacent time periods. This pattern in the residuals is called **autocorrelation**. When a set of data has substantial autocorrelation, the validity of a regression model is in serious doubt.

Residual Plots to Detect Autocorrelation

As mentioned in Section 13.5, one way to detect autocorrelation is to plot the residuals in time order. If a positive autocorrelation effect exists, there will be clusters of residuals with the same sign, and you will readily detect an apparent pattern. If negative autocorrelation exists, residuals will tend to jump back and forth from positive to negative to positive, and so on. Because negative autocorrelation is very rarely seen in regression analysis, the example in this section illustrates positive autocorrelation.

To illustrate positive autocorrelation, consider the case of a package delivery store manager who wants to be able to predict weekly sales. In approaching this problem, the manager has decided to develop a regression model to use the number of customers making purchases as an independent variable. She collects data for a period of 15 weeks and then organizes and stores these data in **FifteenWeeks**). Table 13.4 presents these data.

TABLE 13.4

Customers and Sales
for a Period of 15
Consecutive Weeks

Week	Customers	Sales (\$thousands)	Week	Customers	Sales (\$thousands)
1	794	9.33	9	880	12.07
2	799	8.26	10	905	12.55
3	837	7.48	11	886	11.92
4	855	9.08	12	843	10.27
5	845	9.83	13	904	11.80
6	844	10.09	14	950	12.15
7	863	11.01	15	841	9.64
8	875	11.49			

Because the data are collected over a period of 15 consecutive weeks at the same store, you need to determine whether there is autocorrelation. First, you can develop the simple linear regression model you can use to predict sales based on the number of customers assuming there is no autocorrelation in the residuals. Figure 13.14 presents Excel and Minitab results for these data.

FIGURE 13.14

Excel and Minitab regression results for the Table 13.4 package delivery store data

A	B	C	D	E	F	G
1 Package Delivery Store Sales Analysis						
2						
3 Regression Statistics						
4 Multiple R	0.8108					
5 R Square	0.6574					
6 Adjusted R Square	0.6311					
7 Standard Error	0.9360					
8 Observations	15					
9						
10 ANOVA						
11	df	SS	MS	F	Significance F	
12 Regression	1	21.8604	21.8604	24.9501	0.0002	
13 Residual	13	11.3901	0.8762			
14 Total	14	33.2506				
15						
16	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17 Intercept	-16.0322	5.3102	-3.0192	0.0099	-27.5041	-4.5603
18 Customers	0.0308	0.0062	4.9950	0.0002	0.0175	0.0441

Regression Analysis: Sales versus Customers						
The regression equation is Sales = -16.0 + 0.0308 Customers						
Predictor	Coef	SE Coef	T	P		
Constant	-16.032	5.310	-3.02	0.010		
Customers	0.030760	0.006158	5.00	0.000		
S = 0.936037	R-Sq = 65.7%	R-Sq(adj) = 63.1%				
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	1	21.860	21.860	24.95	0.000	
Residual Error	13	11.390	0.876			
Total	14	33.251				
Durbin-Watson statistic = 0.883003						

From Figure 13.14, observe that r^2 is 0.6574, indicating that 65.74% of the variation in sales is explained by variation in the number of customers. In addition, the Y intercept, b_0 , is -16.0322 and the slope, b_1 , is 0.0308. However, before using this model for prediction, you must perform a residual analysis. Because the data have been collected over a consecutive period of 15 weeks, in addition to checking the linearity, normality, and equal-variance assumptions, you must investigate the independence-of-errors assumption. To do this, you plot the residuals versus time in Figure 13.15 in order to examine whether a pattern in the residuals exists. In Figure 13.15, you can see that the residuals tend to fluctuate up and down in a cyclical pattern. This cyclical pattern provides strong cause for concern about the existence of autocorrelation in the residuals and, therefore, a violation of the independence-of-errors assumption.

FIGURE 13.15

Residual plot for the Table 13.4 package delivery store data



The Durbin-Watson Statistic

The **Durbin-Watson statistic** is used to measure autocorrelation. This statistic measures the correlation between each residual and the residual for the previous time period. Equation (13.15) defines the Durbin-Watson statistic.

DURBIN-WATSON STATISTIC

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} \quad (13.15)$$

where

 e_i = residual at the time period i

In Equation (13.15), the numerator, $\sum_{i=2}^n (e_i - e_{i-1})^2$, represents the squared difference between two successive residuals, summed from the second value to the n th value and the denominator, $\sum_{i=1}^n e_i^2$, represents the sum of the squared residuals. This means that the value of the Durbin-Watson statistic, D , will approach 0 if successive residuals are positively autocorrelated. If the residuals are not correlated, the value of D will be close to 2. (If the residuals are negatively autocorrelated, D will be greater than 2 and could even approach its maximum value of 4.) For the package delivery store data, the Durbin-Watson statistic, D , is 0.8830. (See the Figure 13.16 Excel results or the Figure 13.14 Minitab results.)

FIGURE 13.16

Excel Durbin-Watson statistic worksheet for the package delivery store data

Minitab reports the Durbin-Watson statistic as part of the regression results (see Figure 13.14 on page 512).

	A	B
1	Durbin-Watson Statistic	
2		
3	Sum of Squared Difference of Residuals	10.0575
4	Sum of Squared Residuals	11.3901
5		
6	Durbin-Watson Statistic	0.8830

You need to determine when the autocorrelation is large enough to conclude that there is significant positive autocorrelation. To do so, you compare D to the critical values of the Durbin-Watson statistic found in Table E.8, a portion of which is presented in Table 13.5. The critical values depend on α , the significance level chosen, n , the sample size, and k , the number of independent variables in the model (in simple linear regression, $k = 1$).

TABLE 13.5

Finding Critical Values of the Durbin-Watson Statistic

$\alpha = .05$										
		$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$
n		d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L
15	→	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56
16		1.10	1.37	.98	1.54	.86	1.73	.74	1.93	.62
17		1.13	1.38	1.02	1.54	.90	1.71	.78	1.90	.67
18		1.16	1.39	1.05	1.53	.93	1.69	.82	1.87	.71
										2.06

In Table 13.5, two values are shown for each combination of α (level of significance), n (sample size), and k (number of independent variables in the model). The first value, d_L , represents the lower critical value. If D is below d_L , you conclude that there is evidence of

positive autocorrelation among the residuals. If this occurs, the least-squares method used in this chapter is inappropriate, and you should use alternative methods (see reference 4). The second value, d_U , represents the upper critical value of D , above which you would conclude that there is no evidence of positive autocorrelation among the residuals. If D is between d_L and d_U , you are unable to arrive at a definite conclusion.

For the package delivery store data, with one independent variable ($k = 1$) and 15 values ($n = 15$), $d_L = 1.08$ and $d_U = 1.36$. Because $D = 0.8830 < 1.08$, you conclude that there is positive autocorrelation among the residuals. The least-squares regression analysis of the data shown in Figure 13.14 on page 512 is inappropriate because of the presence of significant positive autocorrelation among the residuals. In other words, the independence-of-errors assumption is invalid. You need to use alternative approaches, discussed in reference 4.

Problems for Section 13.6

LEARNING THE BASICS

- 13.32** The residuals for 10 consecutive time periods are as follows:

Time Period	Residual	Time Period	Residual
1	-5	6	+1
2	-4	7	+2
3	-3	8	+3
4	-2	9	+4
5	-1	10	+5

- a. Plot the residuals over time. What conclusion can you reach about the pattern of the residuals over time?
- b. Based on (a), what conclusion can you reach about the autocorrelation of the residuals?

- 13.33** The residuals for 15 consecutive time periods are as follows:

Time Period	Residual	Time Period	Residual
1	+4	9	+6
2	-6	10	-3
3	-1	11	+1
4	-5	12	+3
5	+2	13	0
6	+5	14	-4
7	-2	15	-7
8	+7		

- a. Plot the residuals over time. What conclusion can you reach about the pattern of the residuals over time?
- b. Compute the Durbin-Watson statistic. At the 0.05 level of significance, is there evidence of positive autocorrelation among the residuals?
- c. Based on (a) and (b), what conclusion can you reach about the autocorrelation of the residuals?

APPLYING THE CONCEPTS

- 13.34** In Problem 13.7 on page 501 concerning the bag sealing equipment at Starbucks, you used the plate gap to predict the tear rating.

- a. Is it necessary to compute the Durbin-Watson statistic in this case? Explain.
- b. Under what circumstances is it necessary to compute the Durbin-Watson statistic before proceeding with the least-squares method of regression analysis?

- 13.35** What is the relationship between the price of crude oil and the price you pay at the pump for gasoline? The file **Oil & Gasoline** contains the price (\$) for a barrel of crude oil (Cushing, Oklahoma, spot price) and a gallon of gasoline (U.S. average conventional spot price) for 181 weeks, ending June 14, 2013. (Data extracted from Energy Information Administration, U.S. Department of Energy, www.eia.doe.gov.)

- a. Construct a scatter plot with the price of oil on the horizontal axis and the price of gasoline on the vertical axis.
- b. Use the least-squares method to develop a simple linear regression equation to predict the price of a gallon of gasoline using the price of a barrel of crude oil as the independent variable.
- c. Interpret the meaning of the slope, b_1 , in this problem.
- d. Plot the residuals versus the time period.
- e. Compute the Durbin-Watson statistic.
- f. At the 0.05 level of significance, is there evidence of positive autocorrelation among the residuals?
- g. Based on the results of (d) through (f), is there reason to question the validity of the model?
- h. What conclusions can you reach concerning the relationship between the price of a barrel of crude oil and the price of a gallon of gasoline?

-  **13.36** A mail-order catalog business that sells personal computer supplies, software, and hardware maintains a centralized warehouse for the distribution of products ordered. Management is currently examining the process of distribution from the warehouse and has the business objective of determining the factors that affect warehouse distribution costs. Currently, a handling fee is added to the order, regardless of the amount of the order. Data that indicate the warehouse distribution

costs and the number of orders received have been collected over the past 24 months and are stored in **Warecost**. The results are:

Months	Distribution Cost (\$thousands)	Number of Orders
1	52.95	4,015
2	71.66	3,806
3	85.58	5,309
4	63.69	4,262
5	72.81	4,296
6	68.44	4,097
7	52.46	3,213
8	70.77	4,809
9	82.03	5,237
10	74.39	4,732
11	70.84	4,413
12	54.08	2,921
13	62.98	3,977
14	72.30	4,428
15	58.99	3,964
16	79.38	4,582
17	94.44	5,582
18	59.74	3,450
19	90.50	5,079
20	93.24	5,735
21	69.33	4,269
22	53.71	3,708
23	89.18	5,387
24	66.80	4,161

- a. Assuming a linear relationship, use the least-squares method to find the regression coefficients b_0 and b_1 .
- b. Predict the monthly warehouse distribution costs when the number of orders is 4,500.
- c. Plot the residuals versus the time period.
- d. Compute the Durbin-Watson statistic. At the 0.05 level of significance, is there evidence of positive autocorrelation among the residuals?
- e. Based on the results of (c) and (d), is there reason to question the validity of the model?
- f. What conclusions can you reach concerning the factors that affect distribution costs?

13.37 A freshly brewed shot of espresso has three distinct components: the heart, body, and crema. The separation of these three

components typically lasts only 10 to 20 seconds. To use the espresso shot in making a latte, a cappuccino, or another drink, the shot must be poured into the beverage during the separation of the heart, body, and crema. If the shot is used after the separation occurs, the drink becomes excessively bitter and acidic, ruining the final drink. Thus, a longer separation time allows the drink-maker more time to pour the shot and ensure that the beverage will meet expectations. An employee at a coffee shop hypothesized that the harder the espresso grounds were tamped down into the portafilter before brewing, the longer the separation time would be. An experiment using 24 observations was conducted to test this relationship. The independent variable Tamp measures the distance, in inches, between the espresso grounds and the top of the portafilter (i.e., the harder the tamp, the greater the distance). The dependent variable Time is the number of seconds the heart, body, and crema are separated (i.e., the amount of time after the shot is poured before it must be used for the customer's beverage). The data are stored in **Espresso**.

- a. Use the least-squares method to develop a simple regression equation with Time as the dependent variable and Tamp as the independent variable.
- b. Predict the separation time for a tamp distance of 0.50 inch.
- c. Plot the residuals versus the time order of experimentation. Are there any noticeable patterns?
- d. Compute the Durbin-Watson statistic. At the 0.05 level of significance, is there evidence of positive autocorrelation among the residuals?
- e. Based on the results of (c) and (d), is there reason to question the validity of the model?
- f. What conclusions can you reach concerning the effect of tamping on the time of separation?

13.38 The owners of a chain of ice cream stores have the business objective of improving the forecast of daily sales so that staffing shortages can be minimized during the summer season. As a starting point, the owners decide to develop a simple linear regression model to predict daily sales based on atmospheric temperature. They select a sample of 21 consecutive days and store the results in **IceCream**. (Hint: Determine which are the independent and dependent variables.)

- a. Assuming a linear relationship, use the least-squares method to compute the regression coefficients b_0 and b_1 .
- b. Predict the sales for a day in which the temperature is 83°F.
- c. Plot the residuals versus the time period.
- d. Compute the Durbin-Watson statistic. At the 0.05 level of significance, is there evidence of positive autocorrelation among the residuals?
- e. Based on the results of (c) and (d), is there reason to question the validity of the model?
- f. What conclusions can you reach concerning the relationship between sales and atmospheric temperature?

13.7 Inferences About the Slope and Correlation Coefficient

In Sections 13.1 through 13.3, regression was used solely for descriptive purposes. You learned how to determine the regression coefficients using the least-squares method and how to predict Y for a given value of X . In addition, you learned how to compute and interpret the standard error of the estimate and the coefficient of determination.

When residual analysis, as discussed in Section 13.5, indicates that the assumptions of a least-squares regression model are not seriously violated and that the straight-line model is appropriate, you can make inferences about the linear relationship between the variables in the population.

t Test for the Slope

To determine the existence of a significant linear relationship between the X and Y variables, you test whether β_1 (the population slope) is equal to 0. The null and alternative hypotheses are as follows:

$$H_0: \beta_1 = 0 \text{ [There is no linear relationship (the slope is zero).]}$$

$$H_1: \beta_1 \neq 0 \text{ [There is a linear relationship (the slope is not zero).]}$$

If you reject the null hypothesis, you conclude that there is evidence of a linear relationship. Equation (13.16) defines the test statistic for the slope, which is based on the sampling distribution of the slope.

TESTING A HYPOTHESIS FOR A POPULATION SLOPE, β_1 , USING THE t TEST

The t_{STAT} test statistic equals the difference between the sample slope and hypothesized value of the population slope divided by S_{b_1} , the standard error of the slope.

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} \quad (13.16)$$

where

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}}$$

$$SSX = \sum_{i=1}^n (X_i - \bar{X})^2$$

The t_{STAT} test statistic follows a t distribution with $n - 2$ degrees of freedom.

Return to the Sunflowers Apparel scenario on page 491. To test whether there is a significant linear relationship between the size of the store and the annual sales at the 0.05 level of significance, refer to the t test results shown in Figure 13.17.

FIGURE 13.17

Excel and Minitab t test for the slope results for the Sunflowers Apparel data

A	B	C	D	E	F	G	Predictor	Coef	SE Coef	T	P
16	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Constant	-1.2088	0.9949	-1.22	0.248
17	Intercept	0.9949	-1.2151	0.2477	-3.3765	0.9588	Profiled Customers	2.0742	0.2536	8.18	0.000
18	Profiled Customers	0.2536	8.1780	0.0000	1.5216	2.6268					

From Figure 13.4 or Figure 13.17,

$$b_1 = +2.0742 \quad n = 14 \quad S_{b_1} = 0.2536$$

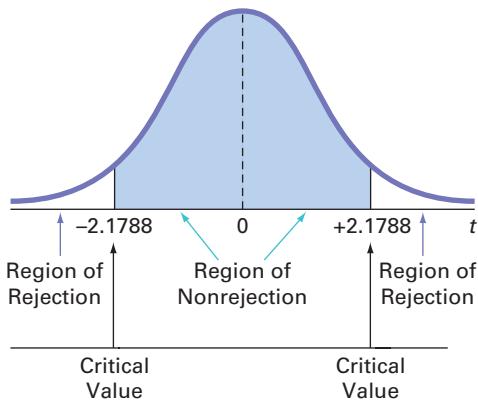
and

$$\begin{aligned} t_{STAT} &= \frac{b_1 - \beta_1}{S_{b_1}} \\ &= \frac{2.0742 - 0}{0.2536} = 8.178 \end{aligned}$$

Using the 0.05 level of significance, the critical value of t with $n - 2 = 12$ degrees of freedom is 2.1788. Because $t_{STAT} = 8.178 > 2.1788$ or because the p -value is 0.0000, which is less than $\alpha = 0.05$, you reject H_0 (see Figure 13.18). Hence, you can conclude that there is a significant linear relationship between mean annual sales and the number of profiled customers.

FIGURE 13.18

Testing a hypothesis about the population slope at the 0.05 level of significance, with 12 degrees of freedom



F Test for the Slope

As an alternative to the t test, in simple linear regression, you can use an F test to determine whether the slope is statistically significant. In Section 10.4, you used the F distribution to test the ratio of two variances. Equation (13.17) defines the F test for the slope as the ratio of the variance that is due to the regression (MSR) divided by the error variance ($MSE = S_{YX}^2$).

TESTING A HYPOTHESIS FOR A POPULATION SLOPE, β_1 , USING THE F TEST

The F_{STAT} test statistic is equal to the regression mean square (MSR) divided by the mean square error (MSE).

$$F_{STAT} = \frac{MSR}{MSE} \quad (13.17)$$

where

$$MSR = \frac{SSR}{1} = SSR$$

$$MSE = \frac{SSE}{n - 2}$$

The F_{STAT} test statistic follows an F distribution with 1 and $n - 2$ degrees of freedom.

Using a level of significance α , the decision rule is

Reject H_0 if $F_{STAT} > F_\alpha$;
otherwise, do not reject H_0 .

Table 13.6 organizes the complete set of results into an analysis of variance (ANOVA) table.

TABLE 13.6

ANOVA Table for Testing the Significance of a Regression Coefficient

Source	df	Sum of Squares	Mean Square (variance)	F
Regression	1	SSR	$MSR = \frac{SSR}{1} = SSR$	$F_{STAT} = \frac{MSR}{MSE}$
Error	$n - 2$	SSE	$MSE = \frac{SSE}{n - 2}$	
Total	$n - 1$	SST		

Figure 13.19, a completed ANOVA table for the Sunflowers sales data (extracted from Figure 13.4), shows that the computed F_{STAT} test statistic is 66.8792 and the p -value is 0.0000.

FIGURE 13.19

Excel and Minitab F test results for the Sunflowers Apparel data

A	B	C	D	E	F
10 ANOVA					
	df	SS	MS	F	Significance F
12 Regression	1	66.7854	66.7854	66.8792	0.0000
13 Residual	12	11.9832	0.9986		
14 Total	13	78.7686			

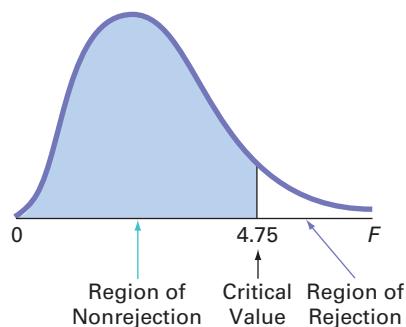
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	66.785	66.785	66.88	0.000
Residual Error	12	11.983	0.999		
Total	13	78.769			

In simple linear regression,
 $t^2 = F$.

Using a level of significance of 0.05, from Table E.5, the critical value of the F distribution, with 1 and 12 degrees of freedom, is 4.75 (see Figure 13.20). Because $F_{STAT} = 66.8792 > 4.75$ or because the p -value = 0.0000 < 0.05, you reject H_0 and conclude that there is a significant linear relationship between the number of profiled customers and annual sales. Because the F test in Equation (13.17) on page 517 is equivalent to the t test in Equation (13.16) on page 516, you reach the same conclusion.

FIGURE 13.20

Regions of rejection and nonrejection when testing for the significance of the slope at the 0.05 level of significance, with 1 and 12 degrees of freedom



Confidence Interval Estimate for the Slope

As an alternative to testing for the existence of a linear relationship between the variables, you can construct a confidence interval estimate of β_1 using Equation (13.18).

CONFIDENCE INTERVAL ESTIMATE OF THE SLOPE, β_1

The confidence interval estimate for the population slope can be constructed by taking the sample slope, b_1 , and adding and subtracting the critical t value multiplied by the standard error of the slope.

$$\begin{aligned} b_1 &\pm t_{\alpha/2}S_{b_1} \\ b_1 - t_{\alpha/2}S_{b_1} &\leq \beta_1 \leq b_1 + t_{\alpha/2}S_{b_1} \end{aligned} \quad (13.18)$$

where

$t_{\alpha/2}$ = critical value corresponding to an upper-tail probability of $\alpha/2$ from the t distribution with $n - 2$ degrees of freedom (i.e., a cumulative area of $1 - \alpha/2$)

From the Figure 13.17 results on page 516,

$$b_1 = 2.0742 \quad n = 14 \quad S_{b_1} = 0.2536$$

To construct a 95% confidence interval estimate, $\alpha/2 = 0.025$, and from Table E.3, $t_{\alpha/2} = 2.1788$. Thus,

$$\begin{aligned} b_1 \pm t_{\alpha/2}S_{b_1} &= 2.0742 \pm (2.1788)(0.2536) \\ &= 2.0742 \pm 0.5526 \\ 1.5216 &\leq \beta_1 \leq 2.6268 \end{aligned}$$

Therefore, you have 95% confidence that the estimated population slope is between 1.5216 and 2.6268. The confidence interval indicates that for each increase of one million profiled customers, predicted annual sales are estimated to increase by at least \$1,521,600 but no more than \$2,626,800. Because both of these values are above 0, you have evidence of a significant linear relationship between annual sales and the number of profiled customers. Had the interval included 0, you would have concluded that there is no evidence of a significant relationship between the variables.

t Test for the Correlation Coefficient

In Section 3.5 on page 133, the strength of the relationship between two numerical variables was measured using the **correlation coefficient**, r . The values of the coefficient of correlation range from -1 for a perfect negative correlation to $+1$ for a perfect positive correlation. You can use the correlation coefficient to determine whether there is a statistically significant linear relationship between X and Y . To do so, you hypothesize that the population correlation coefficient, ρ , is 0. Thus, the null and alternative hypotheses are

$$\begin{aligned} H_0: \rho &= 0 \text{ (no correlation)} \\ H_1: \rho &\neq 0 \text{ (correlation)} \end{aligned}$$

Equation (13.19) defines the test statistic for determining the existence of a significant correlation.

TESTING FOR THE EXISTENCE OF CORRELATION

$$t_{STAT} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} \quad (13.19a)$$

where

$$r = +\sqrt{r^2} \quad \text{if } b_1 > 0$$

$$r = -\sqrt{r^2} \quad \text{if } b_1 < 0$$

The t_{STAT} test statistic follows a t distribution with $n - 2$ degrees of freedom. r is calculated as in Equation (3.17) on page 134:

$$r = \frac{\text{cov}(X, Y)}{S_X S_Y} \quad (13.19b)$$

where

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

$$S_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

$$S_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}}$$

In the Sunflowers Apparel problem, $r^2 = 0.8479$ and $b_1 = +2.0742$ (see Figure 13.4 on page 495). Because $b_1 > 0$, the correlation coefficient for annual sales and store size is the positive square root of r^2 —that is, $r = +\sqrt{0.8479} = +0.9208$. You use Equation (13.19a) to test the null hypothesis that there is no correlation between these two variables. This results in the following t_{STAT} statistic:

$$t_{STAT} = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}} \\ = \frac{0.9208 - 0}{\sqrt{\frac{1 - (0.9208)^2}{14 - 2}}} = 8.178$$

Using the 0.05 level of significance, because $t_{STAT} = 8.178 > 2.1788$, you reject the null hypothesis. You conclude that there is a significant association between annual sales and the number of profiled customers. This t_{STAT} test statistic is equivalent to the t_{STAT} test statistic found when testing whether the population slope, β_1 , is equal to zero.

Problems for Section 13.7

LEARNING THE BASICS

13.39 You are testing the null hypothesis that there is no linear relationship between two variables, X and Y . From your sample of $n = 10$, you determine that $r = 0.80$.

- What is the value of the t test statistic t_{STAT} ?
- At the $\alpha = 0.05$ level of significance, what are the critical values?

- Based on your answers to (a) and (b), what statistical decision should you make?

13.40 You are testing the null hypothesis that there is no linear relationship between two variables, X and Y . From your sample of $n = 18$, you determine that $b_1 = +4.5$ and $S_{b_1} = 1.5$.

- What is the value of t_{STAT} ?
- At the $\alpha = 0.05$ level of significance, what are the critical values?

- c. Based on your answers to (a) and (b), what statistical decision should you make?
- d. Construct a 95% confidence interval estimate of the population slope, β_1 .

13.41 You are testing the null hypothesis that there is no linear relationship between two variables, X and Y . From your sample of $n = 20$, you determine that $SSR = 60$ and $SSE = 40$.

- a. What is the value of F_{STAT} ?
- b. At the $\alpha = 0.05$ level of significance, what is the critical value?
- c. Based on your answers to (a) and (b), what statistical decision should you make?
- d. Compute the correlation coefficient by first computing r^2 and assuming that b_1 is negative.
- e. At the 0.05 level of significance, is there a significant correlation between X and Y ?

APPLYING THE CONCEPTS



13.42 In Problem 13.4 on page 501, you used the percentage of alcohol to predict wine quality. The data are stored in **VinhoVerde**. From the results of that problem, $b_1 = 0.5624$ and $S_{b_1} = 0.1127$.

- a. At the 0.05 level of significance, is there evidence of a linear relationship between the percentage of alcohol and wine quality?
- b. Construct a 95% confidence interval estimate of the population slope, β_1 .

13.43 In Problem 13.5 on page 500, you used the summated rating of a restaurant to predict the cost of a meal. The data are stored in **Restaurants**. Using the results of that problem, $b_1 = 1.4963$ and $S_{b_1} = 0.1379$.

- a. At the 0.05 level of significance, is there evidence of a linear relationship between the summated rating of a restaurant and the cost of a meal?
- b. Construct a 95% confidence interval estimate of the population slope, β_1 .

13.44 In Problem 13.6 on page 501, the owner of a moving company wanted to predict labor hours, based on the number of cubic feet moved. The data are stored in **Moving**. Use the results of that problem.

- a. At the 0.05 level of significance, is there evidence of a linear relationship between the number of cubic feet moved and labor hours?
- b. Construct a 95% confidence interval estimate of the population slope, β_1 .

13.45 In Problem 13.7 on page 501, you used the plate gap in the bag-sealing equipment to predict the tear rating of a bag of coffee. The data are stored in **Starbucks**. Use the results of that problem.

- a. At the 0.05 level of significance, is there evidence of a linear relationship between the plate gap of the bag-sealing machine and the tear rating of a bag of coffee?
- b. Construct a 95% confidence interval estimate of the population slope, β_1 .

13.46 In Problem 13.8 on page 501, you used annual revenues to predict the value of a baseball franchise. The data are stored in **BBRevenue2013**. Use the results of that problem.

- a. At the 0.05 level of significance, is there evidence of a linear relationship between annual revenue and franchise value?
- b. Construct a 95% confidence interval estimate of the population slope, β_1 .

13.47 In Problem 13.9 on page 501, an agent for a real estate company wanted to predict the monthly rent for one-bedroom apartments, based on the size of the apartment. The data are stored in **RentSilverSpring**. Use the results of that problem.

- a. At the 0.05 level of significance, is there evidence of a linear relationship between the size of the apartment and the monthly rent?
- b. Construct a 95% confidence interval estimate of the population slope, β_1 .

13.48 In Problem 13.10 on page 501, you used box office gross to predict DVD revenue. The data are stored in **Movie**. Use the results of that problem.

- a. At the 0.05 level of significance, is there evidence of a linear relationship between box office gross and DVD revenue?
- b. Construct a 95% confidence interval estimate of the population slope, β_1 .

13.49 The volatility of a stock is often measured by its beta value. You can estimate the beta value of a stock by developing a simple linear regression model, using the percentage weekly change in the stock as the dependent variable and the percentage weekly change in a market index as the independent variable. The S&P 500 Index is a common index to use. For example, if you wanted to estimate the beta value for Disney, you could use the following model, which is sometimes referred to as a *market model*:

$$(\text{ % weekly change in Disney }) = \beta_0 + \beta_1 (\text{ % weekly change in S & P 500 index }) + \varepsilon$$

The least-squares regression estimate of the slope b_1 is the estimate of the beta value for Disney. A stock with a beta value of 1.0 tends to move the same as the overall market. A stock with a beta value of 1.5 tends to move 50% more than the overall market, and a stock with a beta value of 0.6 tends to move only 60% as much as the overall market. Stocks with negative beta values tend to move in the opposite direction of the overall market. The following table gives some beta values for some widely held stocks as of August 12, 2012:

Company	Ticker Symbol	Beta
Procter & Gamble	PG	0.32
Dr. Pepper Snapple Group	DPS	-0.02
Disney	DIS	1.07
Apple	AAPL	0.69
eBay	EBAY	0.79
Marriott	MAR	1.32

Source: Data extracted from finance.yahoo.com, June 26, 2013.

- a. For each of the six companies, interpret the beta value.
- b. How can investors use the beta value as a guide for investing?

13.50 Index funds are mutual funds that try to mimic the movement of leading indexes, such as the S&P 500 or the Russell 2000. The beta values (as described in Problem 13.49) for these funds are therefore approximately 1.0, and the estimated market models for these funds are approximately

$$(\text{ % weekly change in index fund }) = 0.0 + 1.0 (\text{ % weekly change in the index })$$

Leveraged index funds are designed to magnify the movement of major indexes. Direxion Funds is a leading provider of leveraged index and other alternative-class mutual fund products for investment advisors and sophisticated investors. Two of the company's funds are shown in the following table:

Name	Ticker Symbol	Description
Daily Small Cap Bull 3x Fund	TNA	300% of the Russell 2000 Index
Monthly S & P Bear 2x Fund	DSSSX	200% of the S&P 500 Index

Source: Data extracted from www.direxionfunds.com.

The estimated market models for these funds are approximately

$$(\% \text{ weekly change in TNA}) = 0.0 + 3.0$$

$$(\% \text{ weekly change in the Russell 2000})$$

$$(\% \text{ weekly change in DSSSX}) = 0.0 + 2.0$$

$$(\% \text{ weekly change in the S & P 500 Index})$$

Thus, if the Russell 2000 Index gains 10% over a period of time, the leveraged mutual fund TNA gains approximately 30%. On the downside, if the same index loses 20%, TNA loses approximately 60%.

- a. The objective of the Direxion Funds Midcap Bull 3x fund, MDU, is 300% of the performance of the S & P Midcap 400 Index. What is its approximate market model?
- b. If the S & P Midcap 400 Index gains 10% in a year, what return do you expect MDU to have?
- c. If the S & P Midcap 400 Index loses 20% in a year, what return do you expect MDU to have?
- d. What type of investors should be attracted to leveraged index funds? What type of investors should stay away from these funds?

13.51 The file **Cereals** contains the calories and sugar, in grams, in one serving of seven breakfast cereals:

Cereal	Calories	Sugar
Kellogg's All Bran	80	6
Kellogg's Corn Flakes	100	2
Wheaties	100	4
Nature's Path Organic Multigrain Flakes	110	4
Kellogg's Rice Krispies	130	4
Post Shredded Wheat Vanilla Almond	190	11
Kellogg's Mini Wheats	200	10

- a. Compute and interpret the coefficient of correlation, r .
- b. At the 0.05 level of significance, is there a significant linear relationship between calories and sugar?

13.52 Movie companies need to predict the gross receipts of an individual movie once the movie has debuted. The following results (stored in **PotterMovies**) are the first weekend gross, the U.S. gross, and the worldwide gross (in \$millions) of the eight Harry Potter movies that debuted from 2001 to 2011:

Title	First Weekend	U.S. Gross	Worldwide Gross
<i>Sorcerer's Stone</i>	90.295	317.558	976.458
<i>Chamber of Secrets</i>	88.357	261.988	878.988
<i>Prisoner of Azkaban</i>	93.687	249.539	795.539
<i>Goblet of Fire</i>	102.335	290.013	896.013
<i>Order of the Phoenix</i>	77.108	292.005	938.469
<i>Half-Blood Prince</i>	77.836	301.460	934.601
<i>Deathly Hallows: Part I</i>	125.017	295.001	955.417
<i>Deathly Hallows: Part II</i>	169.189	381.001	1,328.11

Source: Data extracted from www.the-numbers.com/interactive/comp-Harry-Potter.php.

- a. Compute the coefficient of correlation between first weekend gross and U.S. gross, first weekend gross and worldwide gross, and U.S. gross and worldwide gross.
- b. At the 0.05 level of significance, is there a significant linear relationship between first weekend gross and U.S. gross, first weekend gross and worldwide gross, and U.S. gross and worldwide gross?

13.53 College football is big business, with coaches' salaries, revenues, and expenses in millions of dollars. The file **College Football** contains the coaches' pay and revenue for college football at 105 of the 124 schools that are part of the Division I Football Bowl Subdivision. (Data extracted from "College Football Coaches Continue to See Salary Explosion," *USA Today*, November 20, 2012, p. 8C.)

- a. Compute and interpret the coefficient of correlation, r .
- b. At the 0.05 level of significance, is there a significant linear relationship between a coach's pay and revenue?

13.54 A survey by the Pew Research Center found that social networking is popular in many nations around the world. The file **GlobalSocialMedia** contains the level of social media networking (measured as the percent of individuals polled who use social networking sites) and the GDP per capita based on purchasing power parity (PPP) for each of 25 selected countries. (Data extracted from "Global Digital Communication: Texting, Social Networking Popular Worldwide," The Pew Research Center, updated February 29, 2012, p. 5.)

- a. Compute and interpret the coefficient of correlation, r .
- b. At the 0.05 level of significance, is there a significant linear relationship between GDP and social media usage?
- c. What conclusions can you reach about the relationship between GDP and social media usage?

13.8 Estimation of Mean Values and Prediction of Individual Values

In Chapter 8, you studied the concept of the confidence interval estimate of the population mean. In Example 13.2 on page 497, you used the prediction line to predict the mean value of Y for a given X . The mean annual sales for stores that had four million profiled customers within a fixed radius was predicted to be 7.0879 millions of dollars (\$7,087,900). This estimate, however, is a *point estimate* of the population mean. This section presents methods to develop a confidence interval estimate for the mean response for a given X and for developing a prediction interval for an individual response, Y , for a given value of X .

The Confidence Interval Estimate for the Mean Response

Equation (13.20) defines the **confidence interval estimate for the mean response** for a given X .

CONFIDENCE INTERVAL ESTIMATE FOR THE MEAN OF Y

$$\begin{aligned}\hat{Y}_i &\pm t_{\alpha/2} S_{YX} \sqrt{h_i} \\ \hat{Y}_i - t_{\alpha/2} S_{YX} \sqrt{h_i} &\leq \mu_{Y|X=X_i} \leq \hat{Y}_i + t_{\alpha/2} S_{YX} \sqrt{h_i}\end{aligned}\quad (13.20)$$

where

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX}$$

\hat{Y}_i = predicted value of Y ; $\hat{Y}_i = b_0 + b_1 X_i$

S_{YX} = standard error of the estimate

n = sample size

X_i = given value of X

$\mu_{Y|X=X_i}$ = mean value of Y when $X = X_i$

$$SSX = \sum_{i=1}^n (X_i - \bar{X})^2$$

$t_{\alpha/2}$ = critical value corresponding to an upper-tail probability of $\alpha/2$ from the t distribution with $n - 2$ degrees of freedom (i.e., a cumulative area of $1 - \alpha/2$)

The width of the confidence interval in Equation (13.20) depends on several factors. Increased variation around the prediction line, as measured by the standard error of the estimate, results in a wider interval. As you would expect, increased sample size reduces the width of the interval. In addition, the width of the interval varies at different values of X . When you predict Y for values of X close to \bar{X} , the interval is narrower than for predictions for X values farther away from \bar{X} .

In the Sunflowers Apparel example, suppose you want to construct a 95% confidence interval estimate of the mean annual sales for the entire population of stores that have four million profiled customers ($X = 4$). Using the simple linear regression equation,

$$\begin{aligned}\hat{Y}_i &= -1.2088 + 2.0742X_i \\ &= -1.2088 + 2.0742(4) = 7.0879 \text{ (millions of dollars)}\end{aligned}$$

Also, given the following:

$$\bar{X} = 3.7786 \quad S_{YX} = 0.9993$$

$$SSX = \sum_{i=1}^n (X_i - \bar{X})^2 = 15.5236$$

From Table E.3, $t_{\alpha/2} = 2.1788$. Thus,

$$\hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{h_i}$$

where

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX}$$

so that

$$\begin{aligned} \hat{Y}_i &\pm t_{\alpha/2} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX}} \\ &= 7.0879 \pm (2.1788)(0.9993) \sqrt{\frac{1}{14} + \frac{(4 - 3.7786)^2}{15.5236}} \\ &= 7.0879 \pm 0.5946 \end{aligned}$$

so

$$6.4932 \leq \mu_{Y|X=4} \leq 7.6825$$

Therefore, the 95% confidence interval estimate is that the population mean annual sales are between \$6,493,200 and \$7,682,500 for stores with four million profiled customers.

The Prediction Interval for an Individual Response

In addition to constructing a confidence interval for the mean value of Y , you can also construct a prediction interval for an individual value of Y . Although the form of this interval is similar to that of the confidence interval estimate of Equation (13.20), the prediction interval is predicting an individual value, not estimating a mean. Equation (13.21) defines the **prediction interval for an individual response**, Y , at a given value, X_i , denoted by $Y_{X=X_i}$.

PREDICTION INTERVAL FOR AN INDIVIDUAL RESPONSE, Y

$$\hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{1 + h_i} \quad (13.21)$$

$$\hat{Y}_i - t_{\alpha/2} S_{YX} \sqrt{1 + h_i} \leq Y_{X=X_i} \leq \hat{Y}_i + t_{\alpha/2} S_{YX} \sqrt{1 + h_i}$$

where

$Y_{X=X_i}$ = future value of Y when $X = X_i$

$t_{\alpha/2}$ = critical value corresponding to an upper-tail probability of $\alpha/2$ from the t distribution with $n - 2$ degrees of freedom (i.e., a cumulative area of $1 - \alpha/2$)

In addition, h_i , \hat{Y}_i , S_{YX} , n , and X_i are defined as in Equation (13.20) on page 523.

To construct a 95% prediction interval of the annual sales for an individual store that has four million profiled customers ($X = 4$), you first compute \hat{Y}_i . Using the prediction line:

$$\begin{aligned} \hat{Y}_i &= -1.2088 + 2.0742X_i \\ &= -1.2088 + 2.0742(4) \\ &= 7.0879 \text{ (millions of dollars)} \end{aligned}$$

Also, given the following:

$$\bar{X} = 3.7786 \quad S_{YX} = 0.9993$$

$$SSX = \sum_{i=1}^n (X_i - \bar{X})^2 = 15.5236$$

From Table E.3, $t_{\alpha/2} = 2.1788$. Thus,

$$\hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{1 + h_i}$$

where

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

so that

$$\begin{aligned} \hat{Y}_i &\pm t_{\alpha/2} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX}} \\ &= 7.0879 \pm (2.1788)(0.9993) \sqrt{1 + \frac{1}{14} + \frac{(4 - 3.7786)^2}{15.5236}} \\ &= 7.0879 \pm 2.2570 \end{aligned}$$

so

$$4.8308 \leq Y_{X=4} \leq 9.3449$$

Therefore, with 95% confidence, you predict that the annual sales for an individual store with four million profiled customers is between \$4,830,800 and \$9,344,900.

Figure 13.21 presents Excel and Minitab results for the confidence interval estimate and the prediction interval for the Sunflowers Apparel data. If you compare the results of the confidence interval estimate and the prediction interval, you see that the width of the prediction interval for an individual store is much wider than the confidence interval estimate for the mean. Remember that there is much more variation in predicting an individual value than in estimating a mean value.

FIGURE 13.21

Excel and Minitab confidence interval estimate and prediction interval worksheets for the Sunflowers Apparel data

A	B	
1 Confidence Interval Estimate and Prediction Interval		
2		
3 Data		
4 X Value	4	
5 Confidence Level	95%	
6		
7 Intermediate Calculations		
8 Sample Size	14	=COUNT(SLRData!A:A)
9 Degrees of Freedom	12	=B8 - 2
10 t Value	2.1788	=T.INV.2T(1 - B5, B9)
11 Sample Mean	3.7786	=AVERAGE(SLRData!A:A)
12 Sum of Squared Difference	15.5236	=DEVSQ(SLRData!A:A)
13 Standard Error of the Estimate	0.9993	=COMPUTE!B7
14 h Statistic	0.0746	=1/B8 + (B4 - B11)^2/B12
15 Predicted Y (YHat)	7.0879	=TREND(SLRData!B2:B15, SLRData!A2:A15, B4)
16		
17 For Average Y		
18 Interval Half Width	0.5946	=B10 * B13 * SQRT(B14)
19 Confidence Interval Lower Limit	6.4932	=B15 - B18
20 Confidence Interval Upper Limit	7.6825	=B15 + B18
21		
22 For Individual Response Y		
23 Interval Half Width	2.2570	=B10 * B13 * SQRT(1 + B14)
24 Prediction Interval Lower Limit	4.8308	=B15 - B23
25 Prediction Interval Upper Limit	9.3449	=B15 + B23

Predicted Values for New Observations				
New Obs	Fit	SE Fit	95% CI	95% PI
1	7.088	0.273	(6.493, 7.682)	(4.831, 9.345)

Values of Predictors for New Observations		
Profiled		
New Obs	Customers	
1	4.00	

Problems for Section 13.8

LEARNING THE BASICS

13.55 Based on a sample of $n = 20$, the least-squares method was used to develop the following prediction line: $\hat{Y}_i = 5 + 3X_i$. In addition,

$$S_{YX} = 1.0 \quad \bar{X} = 2 \quad \sum_{i=1}^n (X_i - \bar{X})^2 = 20$$

- a. Construct a 95% confidence interval estimate of the population mean response for $X = 2$.
- b. Construct a 95% prediction interval of an individual response for $X = 2$.

13.56 Based on a sample of $n = 20$, the least-squares method was used to develop the following prediction line: $\hat{Y}_i = 5 + 3X_i$. In addition,

$$S_{YX} = 1.0 \quad \bar{X} = 2 \quad \sum_{i=1}^n (X_i - \bar{X})^2 = 20$$

- a. Construct a 95% confidence interval estimate of the population mean response for $X = 4$.
- b. Construct a 95% prediction interval of an individual response for $X = 4$.
- c. Compare the results of (a) and (b) with those of Problem 13.55 (a) and (b). Which intervals are wider? Why?

APPLYING THE CONCEPTS

13.57 In Problem 13.5 on page 500, you used the summated rating of a restaurant to predict the cost of a meal. The data are stored in **Restaurants**. For these data, $S_{YX} = 9.094$ and $h_i = 0.046319$ when $X = 50$.

- a. Construct a 95% confidence interval estimate of the mean cost of a meal for restaurants that have a summated rating of 50.
- b. Construct a 95% prediction interval of the cost of a meal for an individual restaurant that has a summated rating of 50.
- c. Explain the difference in the results in (a) and (b).

 **SELF Test** **13.58** In Problem 13.4 on page 500, you used the percentage of alcohol to predict wine quality. The data are stored in **VinhoVerde**. For these data, $S_{YX} = 0.9369$ and $h_i = 0.024934$ when $X = 10$.

- a. Construct a 95% confidence interval estimate of the mean wine quality rating for all wines that have 10% alcohol.
- b. Construct a 95% prediction interval of the wine quality rating of an individual wine that has 10% alcohol.
- c. Explain the difference in the results in (a) and (b).

13.59 In Problem 13.7 on page 501, you used the plate gap on the bag-sealing equipment to predict the tear rating of a bag of coffee. The data are stored in **Starbucks**.

- a. Construct a 95% confidence interval estimate of the mean tear rating for all bags of coffee when the plate gap is 0.
- b. Construct a 95% prediction interval of the tear rating for an individual bag of coffee when the plate gap is 0.
- c. Why is the interval in (a) narrower than the interval in (b)?

13.60 In Problem 13.6 on page 501, the owner of a moving company wanted to predict labor hours based on the number of cubic feet moved. The data are stored in **Moving**.

- a. Construct a 95% confidence interval estimate of the mean labor hours for all moves of 500 cubic feet.
- b. Construct a 95% prediction interval of the labor hours of an individual move that has 500 cubic feet.
- c. Why is the interval in (a) narrower than the interval in (b)?

13.61 In Problem 13.9 on page 501, an agent for a real estate company wanted to predict the monthly rent for one-bedroom apartments, based on the size of an apartment. The data are stored in **RentSilverSpring**.

- a. Construct a 95% confidence interval estimate of the mean monthly rental for all one-bedroom apartments that are 800 square feet in size.
- b. Construct a 95% prediction interval of the monthly rental for an individual one-bedroom apartment that is 800 square feet in size.
- c. Explain the difference in the results in (a) and (b).

13.62 In Problem 13.8 on page 501, you predicted the value of a baseball franchise, based on current revenue. The data are stored in **BBRevenue2013**.

- a. Construct a 95% confidence interval estimate of the mean value of all baseball franchises that generate \$250 million of annual revenue.
- b. Construct a 95% prediction interval of the value of an individual baseball franchise that generates \$250 million of annual revenue.
- c. Explain the difference in the results in (a) and (b).

13.63 In Problem 13.10 on page 501, you used box office gross to predict DVD revenue. The data are stored in **Movie**. The company is about to release a movie on DVD that had a box office gross of \$100 million.

- a. What is the predicted DVD revenue?
- b. Which interval is more useful here, the confidence interval estimate of the mean or the prediction interval for an individual response? Explain.
- c. Construct and interpret the interval you selected in (b).

13.9 Potential Pitfalls in Regression

When using regression analysis, some of the potential pitfalls are:

- Lacking awareness of the assumptions of least-squares regression
- Not knowing how to evaluate the assumptions of least-squares regression
- Not knowing what the alternatives are to least-squares regression if a particular assumption is violated
- Using a regression model without knowledge of the subject matter
- Extrapolating outside the relevant range
- Concluding that a significant relationship identified in an observational study is due to a cause-and-effect relationship

The widespread availability of spreadsheet and statistical applications has made regression analysis much more feasible today than it once was. However, many users who have access to such applications do not understand how to use regression analysis properly. Someone who is not familiar with either the assumptions of regression or how to evaluate the assumptions cannot be expected to know what the alternatives to least-squares regression are if a particular assumption is violated.

The data in Table 13.7 (stored in [Anscombe](#)) illustrate the importance of using scatter plots and residual analysis to go beyond the basic number crunching of computing the Y intercept, the slope, and r^2 .

TABLE 13.7

Four Sets of Artificial Data

Data Set A		Data Set B		Data Set C		Data Set D	
X_i	Y_i	X_i	Y_i	X_i	Y_i	X_i	Y_i
10	8.04	10	9.14	10	7.46	8	6.58
14	9.96	14	8.10	14	8.84	8	5.76
5	5.68	5	4.74	5	5.73	8	7.71
8	6.95	8	8.14	8	6.77	8	8.84
9	8.81	9	8.77	9	7.11	8	8.47
12	10.84	12	9.13	12	8.15	8	7.04
4	4.26	4	3.10	4	5.39	8	5.25
7	4.82	7	7.26	7	6.42	19	12.50
11	8.33	11	9.26	11	7.81	8	5.56
13	7.58	13	8.74	13	12.74	8	7.91
6	7.24	6	6.13	6	6.08	8	6.89

Source: Data extracted from F. J. Anscombe, "Graphs in Statistical Analysis," *The American Statistician*, 27 (1973), 17–21.

Anscombe (reference 1) showed that all four data sets given in Table 13.7 have the following identical results:

$$\hat{Y}_i = 3.0 + 0.5X_i$$

$$S_{YX} = 1.237$$

$$S_{b_1} = 0.118$$

$$r^2 = 0.667$$

$$SSR = \text{Explained variation} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = 27.51$$

$$SSE = \text{Unexplained variation} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 13.76$$

$$SST = \text{Total variation} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = 41.27$$

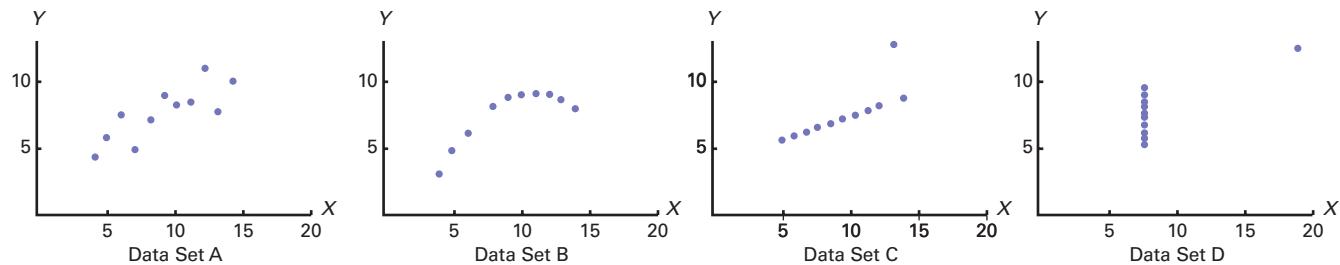
If you stopped the analysis at this point, you would fail to observe the important differences among the four data sets that scatter plots and residual plots can reveal.

From the scatter plots and the residual plots of Figure 13.22, you see how different the data sets are. Each has a different relationship between X and Y . The only data set that seems to approximately follow a straight line is data set A. The residual plot for data set A does not show any obvious patterns or outlying residuals. This is certainly not true for data sets B, C, and D. The scatter plot for data set B shows that a curvilinear regression model is more appropriate. This conclusion is reinforced by the residual plot for data set B. The scatter plot and the residual plot for data set C clearly show an outlying observation. In this case, one approach used is to remove the outlier and reestimate the regression model (see Section 14.8). The scatter plot for data set D represents a situation in which the model is heavily dependent on the outcome of a single data point ($X_8 = 19$ and $Y_8 = 12.50$). Any regression model with this characteristic should be used with caution.

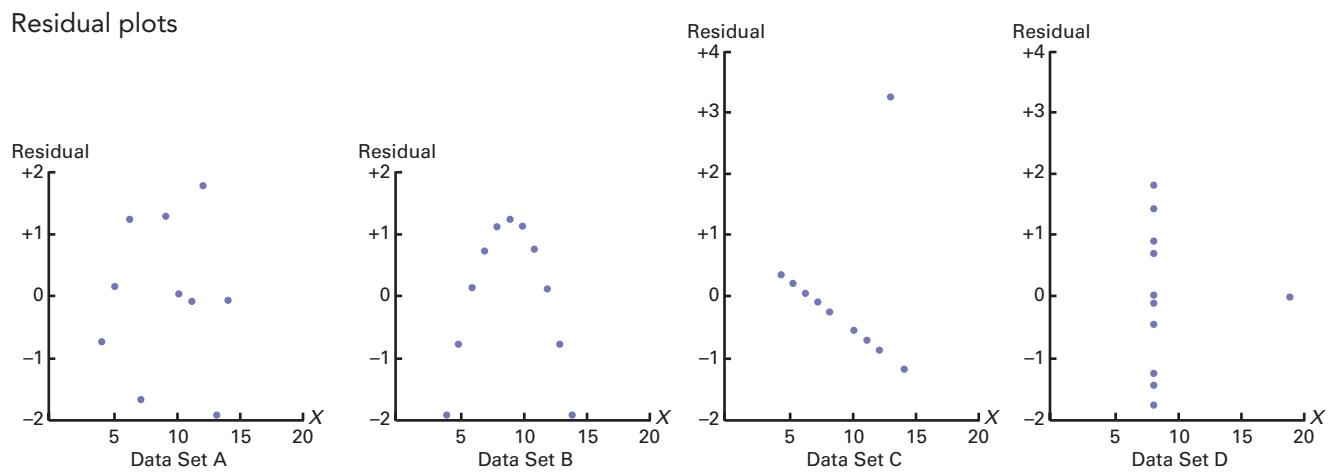
FIGURE 13.22

Scatter plots and residual plots for the data sets A, B, C, and D

Scatter plots



Residual plots



Six Steps for Avoiding the Potential Pitfalls

Constructing scatter plots and residual plots are important tasks in the following six-step strategy to avoid the potential pitfalls in regression. Consider using this strategy *every time* you undertake a regression analysis.

Step 1 Construct a scatter plot to observe the possible relationship between X and Y .

Step 2 Perform a residual analysis to check the assumptions of regression (linearity, independence, normality, equal variance):

- a. Plot the residuals versus the independent variable to determine whether the linear model is appropriate and to check for equal variance.
- b. Construct a histogram, stem-and-leaf display, boxplot, or normal probability plot of the residuals to check for normality.
- c. Plot the residuals versus time to check for independence. (This step is necessary only if the data are collected over time.)

Step 3 If there are violations of the assumptions, use alternative methods to least-squares regression or alternative least-squares models (see reference 4).

Step 4 If there are no violations of the assumptions, carry out tests for the significance of the regression coefficients and develop confidence and prediction intervals.

Step 5 Refrain from making predictions and forecasts outside the relevant range of the independent variable.

Step 6 Remember that the relationships identified in observational studies may or may not be due to cause-and-effect relationships. And while causation implies correlation, correlation does not imply causation.

USING STATISTICS

Knowing Customers at Sunflowers Apparel, Revisited

In the Knowing Customers at Sunflowers Apparel scenario, you were the director of planning for a chain of upscale clothing stores for women. Until now, Sunflowers managers selected sites based on factors such as the availability of a good lease or a subjective opinion that a location seemed like a good place for a store. To make more objective decisions, you used the more systematic DCOVA approach to identify and classify groups of consumers and developed a regression model to analyze the relationship between the number of profiled customers that live within a fixed radius of a Sunflowers store and the annual sales of the store. The

model indicated that about 84.8% of the variation in sales was explained by the number of profiled customers that live within a fixed radius of a Sunflowers store. Furthermore, for each increase of one million profiled customers, mean annual sales were estimated to increase by \$2.0742 million. You can now use your model to help make better decisions when selecting new sites for stores as well as to forecast sales for existing stores.



Fotolia

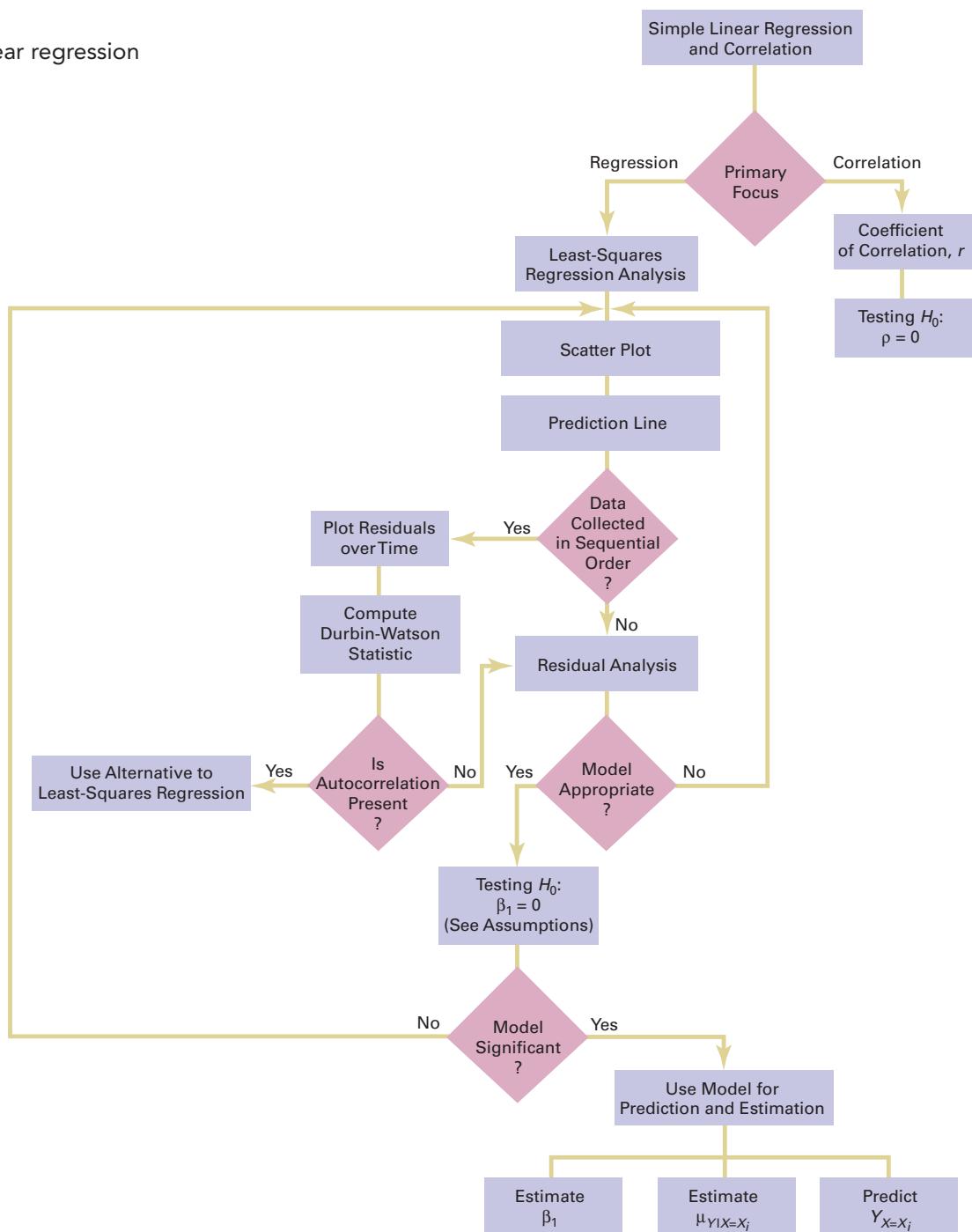
SUMMARY

As you can see from the chapter roadmap in Figure 13.23, this chapter develops the simple linear regression model and discusses the assumptions and how to evaluate them. Once you are assured that the model is appropriate, you can predict values by using the prediction line and test for the

significance of the slope. In Chapters 14 and 15, regression analysis is extended to situations in which more than one independent variable is used to predict the value of a dependent variable.

FIGURE 13.23

Roadmap for simple linear regression



REFERENCES

- Anscombe, F. J. "Graphs in Statistical Analysis." *The American Statistician*, 27(1973): 17–21.
- Hoaglin, D. C., and R. Welsch. "The Hat Matrix in Regression and ANOVA." *The American Statistician*, 32(1978): 17–22.
- Hocking, R. R. "Developments in Linear Regression Methodology: 1959–1982." *Technometrics*, 25(1983): 219–250.
- Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*, 5th ed. New York: McGraw-Hill/Irwin, 2005.
- Microsoft Excel 2013*. Redmond, WA: Microsoft Corp., 2012.
- Minitab Release 16*. State College, PA: Minitab Inc., 2010.
- White, H. "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity." *Econometrica*, 48(1980): 817–838.

KEY EQUATIONS

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (13.1)$$

Simple Linear Regression Equation: The Prediction Line

$$\hat{Y}_i = b_0 + b_1 X_i \quad (13.2)$$

Computational Formula for the Slope, b_1

$$b_1 = \frac{\text{SSXY}}{\text{SSX}} \quad (13.3)$$

Computational Formula for the Y Intercept, b_0

$$b_0 = \bar{Y} - b_1 \bar{X} \quad (13.4)$$

Measures of Variation in Regression

$$SST = SSR + SSE \quad (13.5)$$

Total Sum of Squares (SST)

$$SST = \text{Total sum of squares} = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (13.6)$$

Regression Sum of Squares (SSR)

SSR = Explained variation or regression sum of squares

$$= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad (13.7)$$

Error Sum of Squares (SSE)

SSE = Unexplained variation or error sum of squares

$$= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (13.8)$$

Coefficient of Determination

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST} \quad (13.9)$$

Computational Formula for SST

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} \quad (13.10)$$

Computational Formula for SSR

$$\begin{aligned} SSR &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\ &= b_0 \sum_{i=1}^n Y_i + b_1 \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} \end{aligned} \quad (13.11)$$

Computational Formula for SSE

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i Y_i \quad (13.12)$$

Standard Error of the Estimate

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} \quad (13.13)$$

Residual

$$e_i = Y_i - \hat{Y}_i \quad (13.14)$$

Durbin-Watson Statistic

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} \quad (13.15)$$

Testing a Hypothesis for a Population Slope, β_1 , Using the t Test

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} \quad (13.16)$$

Testing a Hypothesis for a Population Slope, β_1 , Using the F Test

$$F_{STAT} = \frac{MSR}{MSE} \quad (13.17)$$

Confidence Interval Estimate of the Slope, β_1

$$\begin{aligned} b_1 &\pm t_{\alpha/2} S_{b_1} \\ b_1 - t_{\alpha/2} S_{b_1} &\leq \beta_1 \leq b_1 + t_{\alpha/2} S_{b_1} \end{aligned} \quad (13.18)$$

Testing for the Existence of Correlation

$$t_{STAT} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} \quad (13.19a)$$

$$r = \frac{\text{cov}(X, Y)}{S_X S_Y} \quad (13.19b)$$

Confidence Interval Estimate for the Mean of Y

$$\begin{aligned} \hat{Y}_i &\pm t_{\alpha/2} S_{YX} \sqrt{h_i} \\ \hat{Y}_i - t_{\alpha/2} S_{YX} \sqrt{h_i} &\leq \mu_{Y|X=X_i} \leq \hat{Y}_i + t_{\alpha/2} S_{YX} \sqrt{h_i} \end{aligned} \quad (13.20)$$

Prediction Interval for an Individual Response, Y

$$\begin{aligned} \hat{Y}_i &\pm t_{\alpha/2} S_{YX} \sqrt{1 + h_i} \\ \hat{Y}_i - t_{\alpha/2} S_{YX} \sqrt{1 + h_i} &\leq Y_{X=X_i} \leq \hat{Y}_i + t_{\alpha/2} S_{YX} \sqrt{1 + h_i} \end{aligned} \quad (13.21)$$

KEY TERMS

- assumptions of regression 507
 autocorrelation 511
 coefficient of determination 504
 confidence interval estimate for the mean response 523
 correlation coefficient 519
 dependent variable 492
 Durbin-Watson statistic 512
 equal variance 507
 error sum of squares (*SSE*) 502
 explained variation 502
 explanatory variable 493
 homoscedasticity 507
 independence of errors 507
 independent variable 492
 least-squares method 495
 linearity 507
 linear relationship 493
 model 492
 normality 507
 prediction interval for an individual response, \hat{Y} 524
 prediction line 494
 regression analysis 492
 regression coefficient 495
 regression sum of squares (*SSR*) 502
 relevant range 497
 residual 507
 residual analysis 507
 response variable 493
 simple linear regression 492
 simple linear regression equation 494
 slope 494
 standard error of the estimate 505
 total sum of squares (*SST*) 502
 total variation 502
 unexplained variation 502
 Y intercept 494

CHECKING YOUR UNDERSTANDING

- 13.64** What is the interpretation of the Y intercept and the slope in the simple linear regression equation?
- 13.65** What is the interpretation of the coefficient of determination?
- 13.66** When is the unexplained variation (i.e., error sum of squares) equal to 0?
- 13.67** When is the explained variation (i.e., regression sum of squares) equal to 0?
- 13.68** Why should you always carry out a residual analysis as part of a regression model?
- 13.69** What are the assumptions of regression analysis?
- 13.70** How do you evaluate the assumptions of regression analysis?
- 13.71** When and how do you use the Durbin-Watson statistic?
- 13.72** What is the difference between a confidence interval estimate of the mean response, $\mu_{Y|X=X_i}$, and a prediction interval of $\hat{Y}_{X=X_i}$?

CHAPTER REVIEW PROBLEMS

13.73 Can you use Twitter activity to forecast box office receipts on the opening weekend? The following data (stored in [Twitter-Movies](#)) indicate the Twitter activity (“want to see”) and the receipts (\$) per theater on the weekend a movie opened for seven movies:

Movie	Twitter Activity	Receipts (\$)
<i>The Devil Inside</i>	219,509	14,763
<i>The Dictator</i>	6,405	5,796
<i>Paranormal Activity 3</i>	165,128	15,829
<i>The Hunger Games</i>	579,288	36,871
<i>Bridesmaids</i>	6,564	8,995
<i>Red Tails</i>	11,104	7,477
<i>Act of Valor</i>	9,152	8,054

Source: R. Dodes, “Twitter Goes to the Movies,” *The Wall Street Journal*, August 3, 2012, pp. D1–D12.

- Use the least-squares method to compute the regression coefficients b_0 and b_1 .
- Interpret the meaning of b_0 and b_1 in this problem.
- Predict the mean receipts for a movie that has a Twitter activity of 100,000.

- Should you use the model to predict the receipts for a movie that has a Twitter activity of 1,000,000? Why or why not?
- Determine the coefficient of determination, r^2 , and explain its meaning in this problem.
- Perform a residual analysis. Is there any evidence of a pattern in the residuals? Explain.
- At the 0.05 level of significance, is there evidence of a linear relationship between Twitter activity and receipts?
- Construct a 95% confidence interval estimate of the mean receipts for a movie that has a Twitter activity of 100,000 and a 95% prediction interval of the receipts for a single movie that has a Twitter activity of 100,000.
- Based on the results of (a)–(h), do you think that Twitter activity is a useful predictor of receipts on the first weekend a movie opens? What issues about these data might make you hesitant to use Twitter activity to predict receipts?

- 13.74** Management of a soft-drink bottling company has the business objective of developing a method for allocating delivery costs to customers. Although one cost clearly relates to travel time within a particular route, another variable cost reflects the time required to unload the cases of soft drink at the delivery point. To begin, management decided to develop a regression model to predict delivery time based on the number of cases delivered. A sample

of 20 deliveries within a territory was selected. The delivery times and the number of cases delivered were organized in the following table and stored in **Delivery**:

Customer	Delivery		Customer	Delivery	
	Number of Cases	Time (minutes)		Number of Cases	Time (minutes)
1	52	32.1	11	161	43.0
2	64	34.8	12	184	49.4
3	73	36.2	13	202	57.2
4	85	37.8	14	218	56.8
5	95	37.8	15	243	60.6
6	103	39.7	16	254	61.2
7	116	38.5	17	267	58.2
8	121	41.9	18	275	63.1
9	143	44.2	19	287	65.6
10	157	47.1	20	298	67.3

- a. Use the least-squares method to compute the regression coefficients b_0 and b_1 .
- b. Interpret the meaning of b_0 and b_1 in this problem.
- c. Predict the mean delivery time for 150 cases of soft drink.
- d. Should you use the model to predict the delivery time for a customer who is receiving 500 cases of soft drink? Why or why not?
- e. Determine the coefficient of determination, r^2 , and explain its meaning in this problem.
- f. Perform a residual analysis. Is there any evidence of a pattern in the residuals? Explain.
- g. At the 0.05 level of significance, is there evidence of a linear relationship between delivery time and the number of cases delivered?
- h. Construct a 95% confidence interval estimate of the mean delivery time for 150 cases of soft drink and a 95% prediction interval of the delivery time for a single delivery of 150 cases of soft drink.
- i. What conclusions can you reach from (a)–(h) about the relationship between the number of cases and delivery time?

13.75 Measuring the height of a California redwood tree is very difficult because these trees grow to heights of over 300 feet. People familiar with these trees understand that the height of a California redwood tree is related to other characteristics of the tree, including the diameter of the tree at the breast height of a person. The data in **Redwood** represent the height (in feet) and diameter (in inches) at the breast height of a person for a sample of 21 California redwood trees.

- a. Assuming a linear relationship, use the least-squares method to compute the regression coefficients b_0 and b_1 . State the regression equation that predicts the height of a tree based on the tree's diameter at breast height of a person.
- b. Interpret the meaning of the slope in this equation.
- c. Predict the mean height for a tree that has a breast height diameter of 25 inches.
- d. Interpret the meaning of the coefficient of determination in this problem.
- e. Perform a residual analysis on the results and determine the adequacy of the model.

- f. Determine whether there is a significant relationship between the height of redwood trees and the breast height diameter at the 0.05 level of significance.

- g. Construct a 95% confidence interval estimate of the population slope between the height of the redwood trees and breast height diameter.
- h. What conclusions can you reach about the relationship of the diameter of the tree and its height?

13.76 You want to develop a model to predict the assessed value of homes based on their size. A sample of 30 single-family houses listed for sale in Silver Spring, Maryland, a suburb of Washington, DC, is selected to study the relationship between assessed value (in \$thousands) and size (in thousands of square feet), and the data is collected and stored in **SilverSpring**. (Hint: First determine which are the independent and dependent variables.)

- a. Construct a scatter plot and, assuming a linear relationship, use the least-squares method to compute the regression coefficients b_0 and b_1 .
- b. Interpret the meaning of the Y intercept, b_0 , and the slope, b_1 , in this problem.
- c. Use the prediction line developed in (a) to predict the mean assessed value for a house whose size is 2,000 square feet.
- d. Determine the coefficient of determination, r^2 , and interpret its meaning in this problem.
- e. Perform a residual analysis on your results and evaluate the regression assumptions.
- f. At the 0.05 level of significance, is there evidence of a linear relationship between assessed value and size?
- g. Construct a 95% confidence interval estimate of the population slope.
- h. What conclusions can you reach about the relationship between the size of the house and its assessed value?

13.77 You want to develop a model to predict the taxes of houses, based on assessed value. A sample of 30 single-family houses listed for sale in Silver Spring, Maryland, a suburb of Washington, DC, is selected. The taxes (in \$) and the assessed value of the houses (in \$thousands) are recorded and stored in **SilverSpring**. (Hint: First determine which are the independent and dependent variables.)

- a. Construct a scatter plot and, assuming a linear relationship, use the least-squares method to compute the regression coefficients b_0 and b_1 .
- b. Interpret the meaning of the Y intercept, b_0 , and the slope, b_1 , in this problem.
- c. Use the prediction line developed in (a) to predict the mean taxes for a house whose assessed value is \$400,000.
- d. Determine the coefficient of determination, r^2 , and interpret its meaning in this problem.
- e. Perform a residual analysis on your results and evaluate the regression assumptions.
- f. At the 0.05 level of significance, is there evidence of a linear relationship between taxes and assessed value?
- g. What conclusions can you reach concerning the relationship between taxes and assessed value?

13.78 The director of graduate studies at a large college of business has the objective of predicting the grade point average (GPA) of students in an MBA program. The director begins by using the

Graduate Management Admission Test (GMAT) score. A sample of 20 students who have completed two years in the program is selected and stored in **GPIGMAT**.

- a. Construct a scatter plot and, assuming a linear relationship, use the least-squares method to compute the regression coefficients b_0 and b_1 .
- b. Interpret the meaning of the Y intercept, b_0 , and the slope, b_1 , in this problem.
- c. Use the prediction line developed in (a) to predict the mean GPA for a student with a GMAT score of 600.
- d. Determine the coefficient of determination, r^2 , and interpret its meaning in this problem.
- e. Perform a residual analysis on your results and evaluate the regression assumptions.
- f. At the 0.05 level of significance, is there evidence of a linear relationship between GMAT score and GPA?
- g. Construct a 95% confidence interval estimate of the mean GPA of students with a GMAT score of 600 and a 95% prediction interval of the GPA for a particular student with a GMAT score of 600.
- h. Construct a 95% confidence interval estimate of the population slope.
- i. What conclusions can you reach concerning the relationship between GMAT score and GPA?

13.79 An accountant for a large department store has the business objective of developing a model to predict the amount of time it takes to process invoices. Data are collected from the past 32 working days, and the number of invoices processed and completion time (in hours) are stored in **Invoice**. (Hint: First determine which are the independent and dependent variables.)

- a. Assuming a linear relationship, use the least-squares method to compute the regression coefficients b_0 and b_1 .
- b. Interpret the meaning of the Y intercept, b_0 , and the slope, b_1 , in this problem.
- c. Use the prediction line developed in (a) to predict the mean amount of time it would take to process 150 invoices.
- d. Determine the coefficient of determination, r^2 , and interpret its meaning.
- e. Plot the residuals against the number of invoices processed and also against time.
- f. Based on the plots in (e), does the model seem appropriate?
- g. Based on the results in (e) and (f), what conclusions can you reach about the validity of the prediction made in (c)?
- h. What conclusions can you reach about the relationship between the number of invoices and the completion time?

13.80 On January 28, 1986, the space shuttle *Challenger* exploded, and seven astronauts were killed. Prior to the launch, the predicted atmospheric temperature was for freezing weather at the launch site. Engineers for Morton Thiokol (the manufacturer of the rocket motor) prepared charts to make the case that the launch should not take place due to the cold weather. These arguments were rejected, and the launch tragically took place. Upon investigation after the tragedy, experts agreed that the disaster occurred because of leaky rubber O-rings that did not seal properly due to the cold temperature. Data indicating the atmospheric temperature at the time of 23 previous launches and the O-ring damage index are stored in **O-Ring**.

Note: Data from flight 4 is omitted due to unknown O-ring condition.

Sources: Data extracted from *Report of the Presidential Commission on the Space Shuttle Challenger Accident*, Washington, DC, 1986,

Vol. II (H1–H3) and Vol. IV (664); and *Post-Challenger Evaluation of Space Shuttle Risk Assessment and Management*, Washington, DC, 1988, pp. 135–136.

- a. Construct a scatter plot for the seven flights in which there was O-ring damage (O-ring damage index $\neq 0$). What conclusions, if any, can you reach about the relationship between atmospheric temperature and O-ring damage?
- b. Construct a scatter plot for all 23 flights.
- c. Explain any differences in the interpretation of the relationship between atmospheric temperature and O-ring damage in (a) and (b).
- d. Based on the scatter plot in (b), provide reasons why a prediction should not be made for an atmospheric temperature of 31°F, the temperature on the morning of the launch of the *Challenger*.
- e. Although the assumption of a linear relationship may not be valid for the set of 23 flights, fit a simple linear regression model to predict O-ring damage, based on atmospheric temperature.
- f. Include the prediction line found in (e) on the scatter plot developed in (b).
- g. Based on the results in (f), do you think a linear model is appropriate for these data? Explain.
- h. Perform a residual analysis. What conclusions do you reach?

13.81 A baseball analyst would like to study various team statistics for the 2012 baseball season to determine which variables might be useful in predicting the number of wins achieved by teams during the season. He begins by using a team's earned run average (ERA), a measure of pitching performance, to predict the number of wins. He collects the team ERA and team wins for each of the 30 Major League Baseball teams and stores these data in **BB2012**. (Hint: First determine which are the independent and dependent variables.)

- a. Assuming a linear relationship, use the least-squares method to compute the regression coefficients b_0 and b_1 .
- b. Interpret the meaning of the Y intercept, b_0 , and the slope, b_1 , in this problem.
- c. Use the prediction line developed in (a) to predict the mean number of wins for a team with an ERA of 4.50.
- d. Compute the coefficient of determination, r^2 , and interpret its meaning.
- e. Perform a residual analysis on your results and determine the adequacy of the fit of the model.
- f. At the 0.05 level of significance, is there evidence of a linear relationship between the number of wins and the ERA?
- g. Construct a 95% confidence interval estimate of the mean number of wins expected for teams with an ERA of 4.50.
- h. Construct a 95% prediction interval of the number of wins for an individual team that has an ERA of 4.50.
- i. Construct a 95% confidence interval estimate of the population slope.
- j. The 30 teams constitute a population. In order to use statistical inference, as in (f) through (i), the data must be assumed to represent a random sample. What "population" would this sample be drawing conclusions about?
- k. What other independent variables might you consider for inclusion in the model?
- l. What conclusions can you reach concerning the relationship between ERA and wins?

13.82 Can you use the annual revenues generated by National Basketball Association (NBA) franchises to predict franchise values? Figure 2.14 on page 66 shows a scatter plot of revenue with franchise value, and Figure 3.9 on page 136, shows the correlation coefficient. Now, you want to develop a simple linear regression model to predict franchise values based on revenues. (Franchise values and revenues are stored in **NBAValues**.)

- a. Assuming a linear relationship, use the least-squares method to compute the regression coefficients b_0 and b_1 .
- b. Interpret the meaning of the Y intercept, b_0 , and the slope, b_1 , in this problem.
- c. Predict the mean value of an NBA franchise that generates \$150 million of annual revenue.
- d. Compute the coefficient of determination, r^2 , and interpret its meaning.
- e. Perform a residual analysis on your results and evaluate the regression assumptions.
- f. At the 0.05 level of significance, is there evidence of a linear relationship between the annual revenues generated and the value of an NBA franchise?
- g. Construct a 95% confidence interval estimate of the mean value of all NBA franchises that generate \$150 million of annual revenue.
- h. Construct a 95% prediction interval of the value of an individual NBA franchise that generates \$150 million of annual revenue.
- i. Compare the results of (a) through (h) to those of baseball franchises in Problems 13.8, 13.20, 13.30, 13.46, and 13.62 and European soccer teams in Problem 13.83.

13.83 In Problem 13.82 you used annual revenue to develop a model to predict the franchise value of National Basketball Association (NBA) teams. Can you also use the annual revenues generated by European soccer teams to predict franchise values? (European soccer team values and revenues are stored in **SoccerValues2013**.)

- a. Repeat Problem 13.82 (a) through (h) for the European soccer teams.
- b. Compare the results of (a) to those of baseball franchises in Problems 13.8, 13.20, 13.30, 13.46, and 13.62 and NBA franchises in Problem 13.82.

13.84 During the fall harvest season in the United States, pumpkins are sold in large quantities at farm stands. Often, instead of weighing the pumpkins prior to sale, the farm stand operator will just place the pumpkin in the appropriate circular cutout on the counter. When asked why this was done, one farmer replied, "I can tell the weight of the pumpkin from its circumference." To determine whether this was really true, the circumference and weight of each pumpkin from a sample of 23 pumpkins were determined and the results stored in **Pumpkin**.

- a. Assuming a linear relationship, use the least-squares method to compute the regression coefficients b_0 and b_1 .
- b. Interpret the meaning of the slope, b_1 , in this problem.
- c. Predict the mean weight for a pumpkin that is 60 centimeters in circumference.

d. Do you think it is a good idea for the farmer to sell pumpkins by circumference instead of weight? Explain.

e. Determine the coefficient of determination, r^2 , and interpret its meaning.

f. Perform a residual analysis for these data and evaluate the regression assumptions.

g. At the 0.05 level of significance, is there evidence of a linear relationship between the circumference and weight of a pumpkin?

h. Construct a 95% confidence interval estimate of the population slope, β_1 .

13.85 Refer to the discussion of beta values and market models in Problem 13.49 on page 521. The S&P 500 Index tracks the overall movement of the stock market by considering the stock prices of 500 large corporations. The file **StockPrices2012** contains 2012 weekly data for the S&P 500 and three companies. The following variables are included:

WEEK—Week ending on date given

S&P—Weekly closing value for the S&P 500 Index

GE—Weekly closing stock price for General Electric

DISCA—Weekly closing stock price for Discovery Communications

GOOG—Weekly closing stock price for Google

Source: Data extracted from finance.yahoo.com, June 26, 2013.

- a. Estimate the market model for GE. (Hint: Use the percentage change in the S&P 500 Index as the independent variable and the percentage change in GE's stock price as the dependent variable.)
- b. Interpret the beta value for GE.
- c. Repeat (a) and (b) for Discovery Communications.
- d. Repeat (a) and (b) for Google.
- e. Write a brief summary of your findings.

13.86 The file **CEO-Compensation** includes the total compensation (in \$millions) for CEOs of 170 large public companies and the investment return in 2012. (Data extracted from "CEO Pay Rockets as Economy, Stocks Recover," *USA Today*, March 27, 2013, p. 1B.)

- a. Compute the correlation coefficient between compensation and the investment return in 2012.
- b. At the 0.05 level of significance, is the correlation between compensation and the investment return in 2012 statistically significant?
- c. Write a short summary of your findings in (a) and (b). Do the results surprise you?

REPORT WRITING EXERCISE

13.87 In Problems 13.8, 13.20, 13.30, 13.46, 13.62, 13.82, and 13.83, you developed regression models to predict franchise value of major league baseball, NBA basketball, and soccer teams. Now, write a report based on the models you developed. Append to your report all appropriate charts and statistical information.

CASES FOR CHAPTER 13

Managing Ashland MultiComm Services

To ensure that as many trial subscriptions to the *3-For-All* service as possible are converted to regular subscriptions, the marketing department works closely with the customer support department to accomplish a smooth initial process for the trial subscription customers. To assist in this effort, the marketing department needs to accurately forecast the monthly total of new regular subscriptions.

A team consisting of managers from the marketing and customer support departments was convened to develop a better method of forecasting new subscriptions. Previously, after examining new subscription data for the prior three months, a group of three managers would develop a subjective forecast of the number of new subscriptions. Livia Salvador, who was recently hired by the company to provide expertise in quantitative forecasting methods, suggested that the department look for factors that might help in predicting new subscriptions.

Members of the team found that the forecasts in the past year had been particularly inaccurate because in some months, much more time was spent on telemarketing than in other months. Livia collected data (stored in **AMS13**) for the number of new subscriptions and hours spent on telemarketing for each month for the past two years.

1. What criticism can you make concerning the method of forecasting that involved taking the new subscriptions data for the prior three months as the basis for future projections?
2. What factors other than number of telemarketing hours spent might be useful in predicting the number of new subscriptions? Explain.
3.
 - a. Analyze the data and develop a regression model to predict the number of new subscriptions for a month, based on the number of hours spent on telemarketing for new subscriptions.
 - b. If you expect to spend 1,200 hours on telemarketing per month, estimate the number of new subscriptions for the month. Indicate the assumptions on which this prediction is based. Do you think these assumptions are valid? Explain.
 - c. What would be the danger of predicting the number of new subscriptions for a month in which 2,000 hours were spent on telemarketing?

Digital Case

Apply your knowledge of simple linear regression in this Digital Case, which extends the Sunflowers Apparel Using Statistics scenario from this chapter.

Leasing agents from the Triangle Mall Management Corporation have suggested that Sunflowers consider several locations in some of Triangle's newly renovated life-style malls that cater to shoppers with higher-than-mean disposable income. Although the locations are smaller than the typical Sunflowers location, the leasing agents argue that higher-than-mean disposable income in the surrounding community is a better predictor than profiled customers of higher sales. The leasing agents maintain that sample data from 14 Sunflowers stores prove that this is true.

Open **Triangle_Sunflower.pdf** and review the leasing agents' proposal and supporting documents. Then answer the following questions:

1. Should mean disposable income be used to predict sales based on the sample of 14 Sunflowers stores?
2. Should the management of Sunflowers accept the claims of Triangle's leasing agents? Why or why not?
3. Is it possible that the mean disposable income of the surrounding area is not an important factor in leasing new locations? Explain.
4. Are there any other factors not mentioned by the leasing agents that might be relevant to the store leasing decision?

Brynne Packaging

Brynne Packaging is a large packaging company, offering its customers the highest standards in innovative packaging solutions and reliable service. About 25% of the employees at Brynne Packaging are machine operators. The human resources department has suggested that the company

consider using the Wesman Personnel Classification Test (WPCT), a measure of reasoning ability, to screen applicants for the machine operator job. In order to assess the WPCT as a predictor of future job performance, 25 recent applicants were tested using the WPCT; all were hired,

regardless of their WPCT score. At a later time, supervisors were asked to rate the quality of the job performance of these 25 employees, using a 1-to-10 rating scale (where 1 = very low and 10 = very high). Factors considered in the ratings included the employee's output, defect rate, ability to implement continuous quality procedures, and contributions to team problem solving efforts. The file **BrynnePackaging** contains the WPCT scores (WPCT) and job performance ratings (Ratings) for the 25 employees.

1. Assess the significance and importance of WPCT score as a predictor of job performance. Defend your answer.

2. Predict the mean job performance rating for all employees with a WPCT score of 6. Give a point prediction as well as a 95% confidence interval. Do you have any concerns using the regression model for predicting mean job performance rating given the WPCT score of 6?
3. Evaluate whether the assumptions of regression have been seriously violated.

CHAPTER 13 EXCEL GUIDE

EG13.1 TYPES of REGRESSION MODELS

There are no Excel Guide instructions for this section.

EG13.2 DETERMINING the SIMPLE LINEAR REGRESSION EQUATION

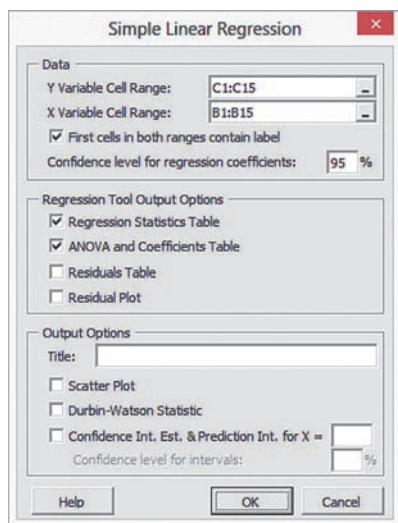
Key Technique Use the **LINEST(cell range of Y variable, cell range of X variable, True, True)** array function to compute the b_1 and b_0 coefficients, the b_1 and b_0 standard errors, r^2 and the standard error of the estimate, the F test statistic and error df , and SSR and SSE .

Example Perform the Figure 13.4 analysis of the Sunflowers Apparel data on page 495.

PHStat Use Simple Linear Regression.

For the example, open to the **DATA worksheet** of the **SiteSelection workbook**. Select **PHStat → Regression → Simple Linear Regression**. In the procedure's dialog box (shown below):

1. Enter C1:C15 as the **Y Variable Cell Range**.
2. Enter B1:B15 as the **X Variable Cell Range**.
3. Check **First cells in both ranges contain label**.
4. Enter **95** as the **Confidence level for regression coefficients**.
5. Check **Regression Statistics Table** and **ANOVA and Coefficients Table**.
6. Enter a **Title** and click **OK**.



The procedure creates a worksheet that contains a copy of your data as well as the worksheet shown in Figure 13.4. For more information about these worksheets, read the following *In-Depth Excel* section.

To create a scatter plot that contains a prediction line and regression equation similar to Figure 13.5 on page 496, modify step 6 by checking **Scatter Plot** before clicking **OK**.

In-Depth Excel Use the **COMPUTE worksheet** of the **Simple Linear Regression workbook** as a template. (Use the **Simple Linear Regression 2007 workbook** if you use an Excel version that is older than Excel 2010.) For the example, the worksheet uses the regression data already in the **SLRDATA worksheet** to perform the regression analysis.

Figure 13.4 does not show the Calculations area in columns K through M. This area contains an array formula in the cell range L2:M6 that contains the expression **LINEST(cell range of Y variable, cell range of X variable, True, True)** to compute the b_1 and b_0 coefficients in cells L2 and M2, the b_1 and b_0 standard errors in cells L3 and M3, r^2 and the standard error of the estimate in cells L4 and M4, the F test statistic and error df in cells L5 and M5, and SSR and SSE in cells L6 and M6. In cell L9, the expression **T.INV.2T(1 - confidence level, Error degrees of freedom)** computes the critical value for the t test. Open the **COMPUTE_FORMULAS worksheet** to examine all the formulas in the worksheet, some of which are discussed in later sections in this Excel Guide.

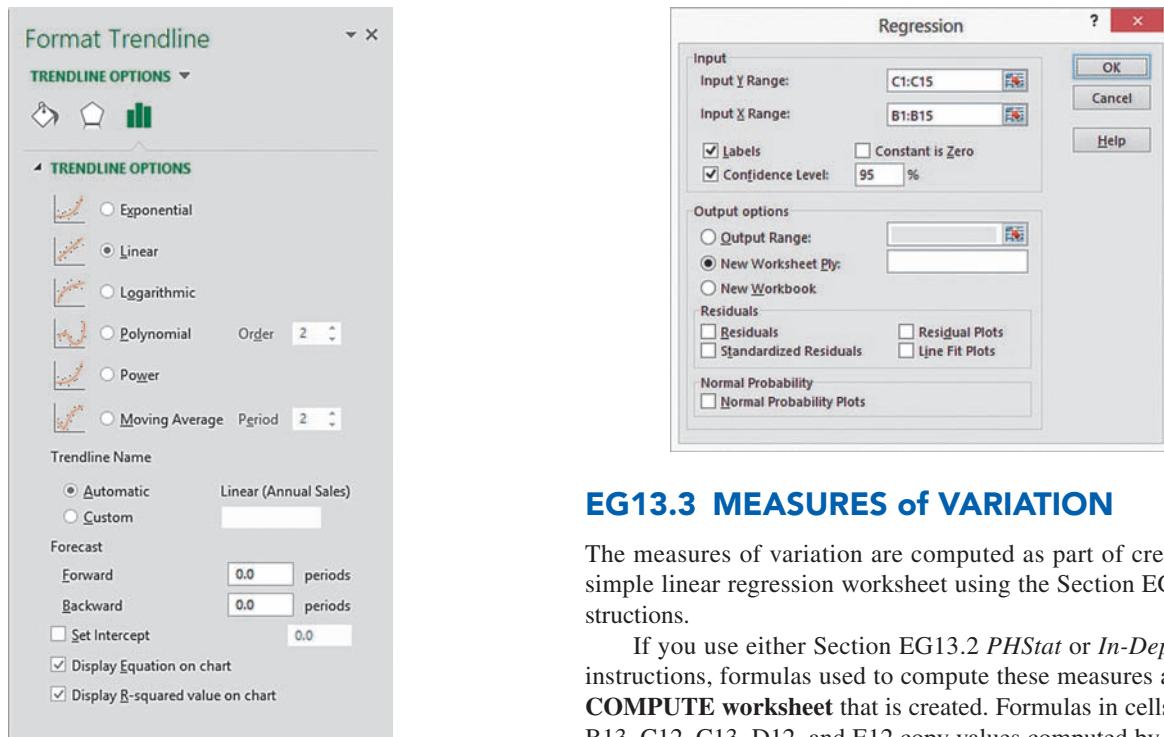
To perform simple linear regression for other data, paste the regression data into the **SLRDATA worksheet**. Paste the values for the **X** variable into column A and the values for the **Y** variable into column B. Then, open to the **COMPUTE worksheet**. Enter the confidence level in cell L8 and edit the array formula in the cell range L2:M6. To edit the array formula, first select L2:M6, next make changes to the array formula, and then, while holding down the **Control** and **Shift** keys (or the **Command** key on a Mac), press the **Enter** key.

To create a scatter plot that contains a prediction line and regression equation similar to Figure 13.5 on page 496, first use the Section EG2.5 *In-Depth Excel* scatter plot instructions with the Table 13.1 Sunflowers Apparel data to create a scatter plot. Then select the chart and:

1. Select **Design → Add Chart Element → Trendline → More Trendline Options**.

In the Format Trendline pane (shown on page 539):

2. Click **Linear**, check both **Display Equation on chart** and **Display R-squared value on chart**.



If you use an Excel version that is older than Excel 2010, after selecting the chart:

1. Select Layout → Trendline → More Trendline Options.

In the Format Trendline dialog box (similar to the Format Trendline pane):

2. Click Trendline Options in the left pane. In the Trendline Options right pane, click Linear, check Display Equation on chart, check Display R-squared value on chart, and then click Close.

For scatter plots of other data, if the **X** axis does not appear at the bottom of the plot, right-click the **Y axis** and click **Format Axis** from the shortcut menu. In the Format Axis dialog box, click **Axis Options** in the left pane. In the **Axis Options** pane on the right, click **Axis value** and in its box enter the value shown in the dimmed **Minimum** box at the top of the pane. Then click **Close**.

Analysis ToolPak Use Regression.

For the example, open to the **DATA worksheet** of the **SiteSelection workbook** and:

1. Select Data → Data Analysis.

2. In the Data Analysis dialog box, select **Regression** from the **Analysis Tools** list and then click **OK**.

In the Regression dialog box (shown in next column):

3. Enter **C1:C15** as the **Input Y Range** and enter **B1:B15** as the **Input X Range**.

4. Check **Labels** and check **Confidence Level** and enter **95** in its box.

5. Click **New Worksheet Ply** and then click **OK**.

EG13.3 MEASURES of VARIATION

The measures of variation are computed as part of creating the simple linear regression worksheet using the Section EG13.2 instructions.

If you use either Section EG13.2 **PHStat** or **In-Depth Excel** instructions, formulas used to compute these measures are in the **COMPUTE worksheet** that is created. Formulas in cells B5, B7, B13, C12, C13, D12, and E12 copy values computed by the array formula in cell range L2:M6. In cell F12, the **F.DIST.RT(F test statistic, regression degrees of freedom, error degrees of freedom)**, function computes the *p*-value for the *F* test for the slope, discussed in Section 13.7. (The similar FDIST function is used in the COMPUTE worksheet of the Simple Linear Regression 2007 workbook.)

EG13.4 ASSUMPTIONS of REGRESSION

There are no Excel Guide instructions for this section.

EG13.5 RESIDUAL ANALYSIS

Key Technique Use arithmetic formulas to compute the residuals. To evaluate assumptions, use the Section EG2.5 scatter plot instructions for constructing residual plots and the Section EG6.3 instructions for constructing normal probability plots.

Example Compute the residuals for the Table 13.1 Sunflowers Apparel data on page 494.

PHStat Use the Section EG13.2 **PHStat** instructions. Modify step 5 by checking **Residuals Table** and **Residual Plot** in addition to checking **Regression Statistics Table** and **ANOVA and Coefficients Table**. To construct a normal probability plot, follow the Section EG6.3 **PHStat** instructions using the cell range of the residuals as the **Variable Cell Range** in step 1.

In-Depth Excel Use the **RESIDUALS worksheet** of the **Simple Linear Regression workbook** as a template.

This worksheet already computes the residuals for the example. Column C formulas compute the predicted *Y* values (labeled Predicted Annual Sales in Figure 13.10 on page 508) by first multiplying the *X* values by the b_1 coefficient in cell B18 of the COMPUTE worksheet and then adding the b_0 coefficient (in

cell B17 of COMPUTE). Column E formulas compute residuals by subtracting the predicted Y values from the Y values (labeled Annual Sales in Figure 13.10).

For other problems, modify this worksheet by pasting the X values into column B and the Y values into column D. Then, for sample sizes smaller than 14, delete the extra rows. For sample sizes greater than 14, copy the column C and E formulas down through the row containing the last pair and X and Y values and add the new observation numbers in column A.

To construct a residual plot similar to Figure 13.11 on page 509, use the original X variable and the residuals (plotted as the Y variable) as the chart data and follow the Section EG2.5 scatter plot instructions. To construct a normal probability plot, follow the Section EG6.3 *In-Depth Excel* instructions, using the cell range of the residuals as the **Variable Cell Range**.

Analysis ToolPak Use the Section EG13.2 *Analysis ToolPak* instructions.

Modify step 5 by checking **Residuals** and **Residual Plots** before clicking **New Worksheet Ply** and then **OK**. To construct a residual plot or normal probability plot, use the *In-Depth Excel* instructions.

EG13.6 MEASURING AUTOCORRELATION: the DURBIN-WATSON STATISTIC

Key Technique Use the **SUMXMY2(cell range of the second through last residual, cell range of the first through the second-to-last residual)** function to compute the sum of squared difference of the residuals, the numerator in Equation (13.15) on page 513, and use the **SUMSQ(cell range of the residuals)** function to compute the sum of squared residuals, the denominator in Equation (13.15).

Example Compute the Durbin-Watson statistic for the Sunflowers Apparel data on page 494.

PHStat Use the *PHStat* instructions at the beginning of Section EG13.2. Modify step 6 by checking the **Durbin-Watson Statistic** output option before clicking **OK**.

In-Depth Excel Use the **DURBIN_WATSON worksheet** of the **Simple Linear Regression** workbook as a template. The worksheet uses the **SUMXMY2** function in cell B3 and the **SUMSQ** function in cell B4.

The **DURBIN_WATSON worksheet** of the **Package Delivery workbook** computes the statistic for the Figure 13.16 package delivery store example on page 513. (This workbook also uses the COMPUTE and RESIDUALS worksheet templates from the Simple Linear Regression workbook.)

To compute the Durbin-Watson statistic for other problems, first create the simple linear regression model and the residuals for the problem, using the Sections EG13.2 and EG13.5 *In-Depth Excel* instructions. Then open the **DURBIN_WATSON worksheet** and edit the formulas in cell B3 and B4 to point to the proper cell ranges of the new residuals.

EG13.7 INFERENCEs ABOUT the SLOPE and CORRELATION COEFFICIENT

The t test for the slope and F test for the slope are included in the worksheet created by using the Section EG13.2 instructions. The t test computations in the worksheets created by using the *PHStat* and *In-Depth Excel* instructions are discussed in Section EG13.2. The F test computations are discussed in Section EG13.3.

EG13.8 ESTIMATION of MEAN VALUES and PREDICTION of INDIVIDUAL VALUES

Key Technique Use the **TREND(Y variable cell range, X variable cell range, X value)** function to compute the predicted Y value for the X value and use the **DEVSQ(X variable cell range)** function to compute the SSX value.

Example Compute the Figure 13.21 confidence interval estimate and prediction interval for the Sunflowers Apparel data that is shown on page 494.

PHStat Use the Section EG13.2 *PHStat* instructions but replace step 6 with these steps 6 and 7:

1. Check **Confidence Int. Est. & Prediction Int. for $X =$** and enter **4** in its box. Enter **95** as the percentage for **Confidence level for intervals**.
2. Enter a **Title** and click **OK**.

The additional worksheet created is discussed in the following *In-Depth Excel* instructions.

In-Depth Excel Use the **CIEandPI worksheet** of the **Simple Linear Regression** workbook, as a template.

The worksheet already contains the data and formulas for the example. The worksheet uses the **T.INV.2T(1 - confidence level, degrees of freedom)** function to compute the t critical value in cell B10 and the TREND function to compute the predicted Y value for the X value in cell B15. In cell B12, the function **DEVSQ(SLRData!A:A)** computes the SSX value that is used, in turn, to help compute the h statistic in cell B14.

To compute a confidence interval estimate and prediction interval for other problems:

1. Paste the regression data into the **SLRData worksheet**. Use column A for the X variable data and column B for the Y variable data.
2. Open to the **CIEandPI worksheet**.

In the **CIEandPI worksheet**:

3. Change values for the **X Value** and **Confidence Level**, as is necessary.
4. Edit the cell ranges used in the cell B15 formula that uses the TREND function to refer to the new cell ranges for the Y and X variables.

CHAPTER 13 MINITAB GUIDE

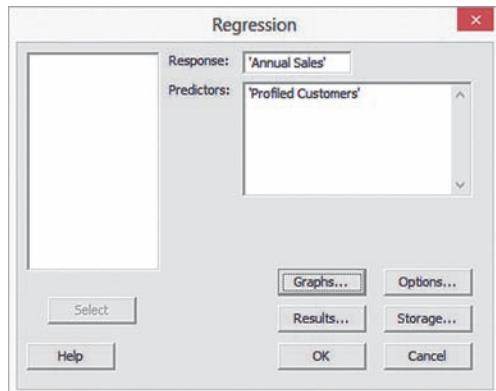
MG13.1 TYPES of REGRESSION MODELS

There are no Minitab Guide instructions for this section.

MG13.2 DETERMINING the SIMPLE LINEAR REGRESSION EQUATION

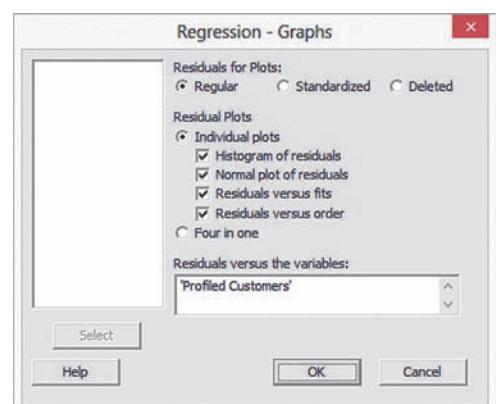
Use **Regression** to perform a simple linear regression analysis. For example, to perform the Figure 13.4 analysis of the Sunflowers Apparel data on page 495, open to the **SiteSelection worksheet**. Select **Stat → Regression → Regression**. In the Regression dialog box (shown below):

1. Double-click **C3 Annual Sales** in the variables list to add 'Annual Sales' to the **Response** box.
2. Double-click **C2 Profiled Customers** in the variables list to add 'Profiled Customers' to the **Predictors** box.
3. Click **Graphs**.



In the Regression - Graphs dialog box (shown below):

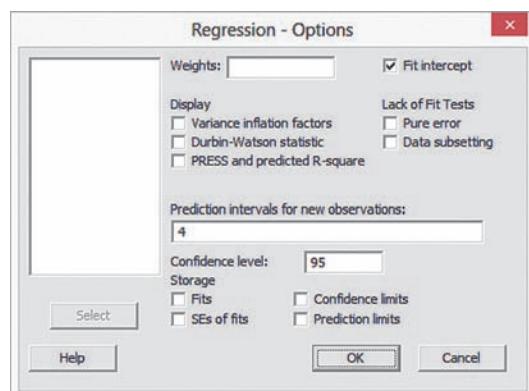
4. Click **Regular** (in Residuals for Plots) and **Individual Plots** (in Residual Plots).
5. Check **Histogram of residuals**, **Normal plot of residuals**, **Residuals versus fits**, and **Residuals versus order** and then press **Tab**.
6. Double-click **C2 Profiled Customers** in the variables list to add 'Profiled Customers' in the **Residuals versus the variables** box.
7. Click **OK**.



8. Back in the Regression dialog box, click **Results**.

In the Regression - Results dialog box (not shown):

9. Click **Regression equation, table of coefficients, s, R-squared, and basic analysis of variance** and then click **OK**.
 10. Back in the Regression dialog box, click **Options**.
- In the Regression - Options dialog box (shown below):
11. Check **Fit Intercept**.
 12. Clear all the **Display** and **Lack of Fit Test** check boxes.
 13. Enter **4** in the **Prediction intervals for new observations** box.
 14. Enter **95** in the **Confidence level** box.
 15. Click **OK**.



16. Back in the Regression dialog box, click **OK**.

To create a scatter plot that contains a prediction line and regression equation similar to Figure 13.5 on page 496, use the Section MG2.6 scatter plot instructions with the Table 13.1 Sunflowers Apparel data.

MG13.3 MEASURES of VARIATION

The measures of variation are computed in the Analysis of Variance table that is part of the simple linear regression results created using the Section MG13.2 instructions.

MG13.4 ASSUMPTIONS

There are no Minitab Guide instructions for this section.

MG13.5 RESIDUAL ANALYSIS

Selections in step 5 of the Section MG13.2 instructions create the residual plots and normal probability plots necessary for residual analysis. To create the list of residual values similar to the last column in Figure 13.10 on page 508, replace step 15 of the Section MG13.2 instructions with these steps 15 through 17:

15. Click **Storage**.
16. In the Regression - Storage dialog box, check **Residuals** and then click **OK**.
17. Back in the Regression dialog box, click **OK**.

MG13.6 MEASURING AUTOCORRELATION: THE DURBIN-WATSON STATISTIC

To compute the Durbin-Watson statistic, use the Section MG13.2 instructions but check **Durbin-Watson statistic** (in the Regression - Options dialog box) as part of step 12.

MG13.7 INFERENCES ABOUT the SLOPE AND CORRELATION COEFFICIENT

The *t* test for the slope and *F* test for the slope are included in the results created by using the Section MG13.2 instructions.

MG13.8 ESTIMATION of MEAN VALUES and PREDICTION of INDIVIDUAL VALUES

The confidence interval estimate and prediction interval are included in the results created by using the Section MG13.2 instructions.

CHAPTER 14

CONTENTS

- 14.1 Developing a Multiple Regression Model
- 14.2 r^2 , Adjusted r^2 , and the Overall F Test
- 14.3 Residual Analysis for the Multiple Regression Model
- 14.4 Inferences Concerning the Population Regression Coefficients
- 14.5 Testing Portions of the Multiple Regression Model
- 14.6 Using Dummy Variables and Interaction Terms in Regression Models
- 14.7 Logistic Regression
- 14.8 Influence Analysis

USING STATISTICS: The Multiple Effects of OmniPower Bars, Revisited

CHAPTER 14 EXCEL GUIDE

CHAPTER 14 MINITAB GUIDE

OBJECTIVES

- To learn to develop a multiple regression model
- To learn to interpret the regression coefficients
- To learn to determine which independent variables to include in the regression model
- To learn to determine which independent variables are most important in predicting a dependent variable
- To learn to use categorical independent variables in a regression model
- To use logistic regression to predict a categorical dependent variable
- To learn to identify individual observations that may be unduly influencing the multiple regression model

Introduction to Multiple Regression

USING STATISTICS

The Multiple Effects of OmniPower Bars

You are a marketing manager for OmniFoods, with oversight for nutrition bars and similar snack items. You seek to revive the sales of OmniPower, the company's primary product in this category. Originally marketed as a high-energy bar to runners, mountain climbers, and other athletes, OmniPower reached its greatest sales in an earlier time, when high-energy bars were most popular with consumers. Now, you seek to remarket the product as a nutrition bar to benefit from the booming market for such bars.

Because the marketplace already contains several successful nutrition bars, you need to develop an effective marketing strategy. In particular, you need to determine the effect that price and in-store promotional expenses (special in-store coupons, signs, and displays as well as the cost of free samples) will have on sales of OmniPower. Before marketing the bar nationwide, you plan to conduct a test-market study of OmniPower sales, using a sample of 34 stores in a supermarket chain.

How can you extend the linear regression methods discussed in Chapter 13 to incorporate the effects of price *and* promotion into the same model? How can you use this model to improve the success of the nationwide introduction of OmniPower?



Ariwasabi/Shutterstock

Chapter 13 discusses simple linear regression models that use *one* numerical independent variable, X , to predict the value of a numerical dependent variable, Y . Often you can make better predictions by using *more than one* independent variable. This chapter introduces you to **multiple regression models** that use two or more independent variables to predict the value of a dependent variable.

14.1 Developing a Multiple Regression Model

In the OmniPower Bars scenario, your business objective, to determine the effect that price and in-store promotional expenses will have on sales, calls for examining a multiple regression model in which the price of an OmniPower bar in cents (X_1) and the monthly budget for in-store promotional expenditures in dollars (X_2) are the independent variables and the number of OmniPower bars sold in a month (Y) is the dependent variable.

To develop this model, you collect data from a sample of 34 stores in a supermarket chain selected for a test-market study of OmniPower. You choose stores in a way to ensure that they all have approximately the same monthly sales volume. You organize and store the data collected in **OmniPower**. Table 14.1 presents these data.

TABLE 14.1

Monthly OmniPower Sales, Price, and Promotional Expenditures

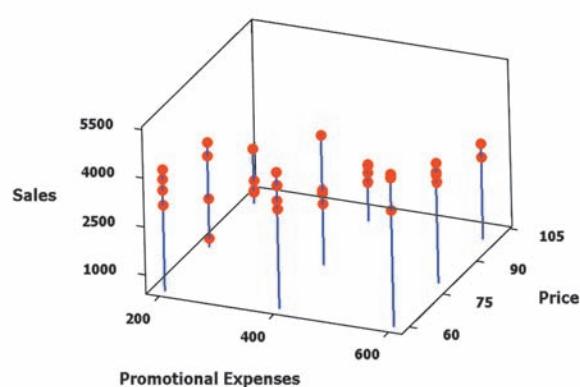
Store	Sales	Price	Promotion	Store	Sales	Price	Promotion
1	4,141	59	200	18	2,730	79	400
2	3,842	59	200	19	2,618	79	400
3	3,056	59	200	20	4,421	79	400
4	3,519	59	200	21	4,113	79	600
5	4,226	59	400	22	3,746	79	600
6	4,630	59	400	23	3,532	79	600
7	3,507	59	400	24	3,825	79	600
8	3,754	59	400	25	1,096	99	200
9	5,000	59	600	26	761	99	200
10	5,120	59	600	27	2,088	99	200
11	4,011	59	600	28	820	99	200
12	5,015	59	600	29	2,114	99	400
13	1,916	79	200	30	1,882	99	400
14	675	79	200	31	2,159	99	400
15	3,636	79	200	32	1,602	99	400
16	3,224	79	200	33	3,354	99	600
17	2,295	79	400	34	2,927	99	600

When there are two independent variables in the multiple regression model, using a three-dimensional (3D) scatter plot can help suggest a starting point for analysis. Figure 14.1 on page 545 presents a 3D scatter plot of the OmniPower data. In this figure, points are plotted at a height equal to their sales and have drop lines down to their corresponding price and promotional expense values. Rotating 3D plots can sometimes reveal patterns. One rotated view (Figure 14.1 right) suggests a negative linear relationship between sales and price (sales decrease as price increases) and a positive linear relationship between sales and promotional expenses (sales increase as those expenses increase). These relationships are not easily seen in the original orientation of the scatter plot.

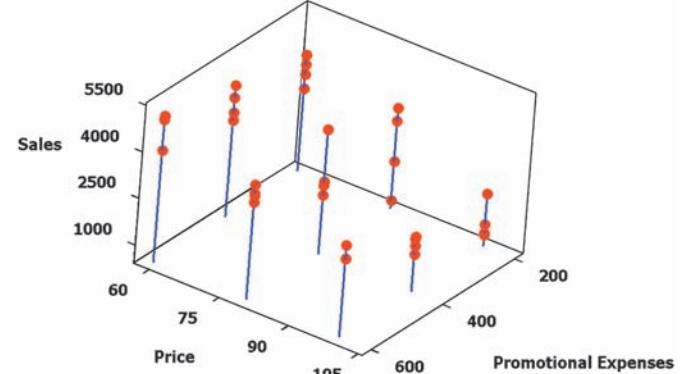
FIGURE 14.1

Original (left) and rotated (right) Minitab 3D scatter plot of the monthly OmniPower sales, price, and promotional expenses

3D Scatterplot of Sales vs Price vs Promotional Expenses



Excel does not include the capability to construct 3D scatter plots



Interpreting the Regression Coefficients

When there are several independent variables, you can extend the simple linear regression model of Equation (13.1) on page 493 by assuming a linear relationship between each independent variable and the dependent variable. For example, with k independent variables, the multiple regression model is expressed in Equation (14.1).

MULTIPLE REGRESSION MODEL WITH k INDEPENDENT VARIABLES

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \varepsilon_i \quad (14.1)$$

where

β_0 = Y intercept

β_1 = slope of Y with variable X_1 , holding variables X_2, X_3, \dots, X_k constant

β_2 = slope of Y with variable X_2 , holding variables X_1, X_3, \dots, X_k constant

β_3 = slope of Y with variable X_3 , holding variables X_1, X_2, \dots, X_k constant

\vdots

β_k = slope of Y with variable X_k holding variables $X_1, X_2, X_3, \dots, X_{k-1}$ constant

ε_i = random error in Y for observation i

Equation (14.2) defines the multiple regression model with two independent variables.

MULTIPLE REGRESSION MODEL WITH TWO INDEPENDENT VARIABLES

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (14.2)$$

where

β_0 = intercept

β_1 = slope of Y with variable X_1 , holding variable X_2 constant

β_2 = slope of Y with variable X_2 , holding variable X_1 constant

ε_i = random error in Y for observation i

Compare the multiple regression model to the simple linear regression model [Equation (13.1) on page 493]:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

In the simple linear regression model, the slope, β_1 , represents the change in the mean of Y per unit change in X and does not take into account any other variables. In the multiple regression model with two independent variables [Equation (14.2)], the slope, β_1 , represents the change in the mean of Y per unit change in X_1 , taking into account the effect of X_2 .

As in the case of simple linear regression, you use the least-squares method to compute the sample regression coefficients b_0 , b_1 , and b_2 as estimates of the population parameters β_0 , β_1 , and β_2 . Equation (14.3) defines the regression equation for a multiple regression model with two independent variables.

Student Tip

Because multiple regression computations are more complex than computations for simple linear regression, always use a computerized method to obtain multiple regression results.

MULTIPLE REGRESSION EQUATION WITH TWO INDEPENDENT VARIABLES

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} \quad (14.3)$$

Figure 14.2 shows Excel and Minitab results for the OmniPower sales data multiple regression model. In these results, the b_0 coefficient is labeled Intercept by Excel and Constant by Minitab.

FIGURE 14.2

Excel and Minitab results for the OmniPower sales multiple regression model

A	B	C	D	E	F	G
OmniPower Sales Multiple Regression Model						
Regression Statistics						
Multiple R	0.8705					
R Square	0.7577					
Adjusted R Square	0.7421					
Standard Error	638.0653					
Observations	34					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	2	39472730.7730	1973635.3865	48.4771	0.0000	
Residual	31	12620946.6682	407127.3119			
Total	33	52093677.4412				
Coefficients						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	5837.5208	628.1502	9.2932	0.0000	4556.3999	7118.6416
Price	-53.2173	6.8522	-7.7664	0.0000	-67.1925	-39.2421
Promotional Expenses	3.6131	0.6852	5.2728	0.0000	2.2155	5.0106

Regression Analysis: Sales versus Price, Promotion

The regression equation is
Sales = 5838 - 53.2 Price + 3.61 Promotion

Predictor	Coeff	SE Coef	T	P
Constant	5837.5	628.2	9.29	0.000
Price	-53.217	6.852	-7.77	0.000
Promotion	3.6131	0.6852	5.27	0.000

S = 638.065 R-Sq = 75.8% R-Sq(adj) = 74.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	39472731	19736365	48.48	0.000
Residual Error	31	12620947	407127		
Total	33	52093677			

From Figure 14.2, the computed values of the three regression coefficients are

$$b_0 = 5,837.5208 \quad b_1 = -53.2173 \quad b_2 = 3.6131$$

Therefore, the multiple regression equation is

$$\hat{Y}_i = 5,837.5208 - 53.2173X_{1i} + 3.6131X_{2i}$$

where

$$\begin{aligned}\hat{Y}_i &= \text{predicted monthly sales of OmniPower bars for store } i \\ X_{1i} &= \text{price of OmniPower bar (in cents) for store } i \\ X_{2i} &= \text{monthly in-store promotional expenditures (in \$) for store } i\end{aligned}$$

The sample Y intercept ($b_0 = 5,837.5208$) estimates the number of OmniPower bars sold in a month if the price is \$0.00 and the total amount spent on promotional expenditures is also \$0.00. Because these values of price and promotion are outside the range of price and promotion used in the test-market study, and because they make no sense in the context of the problem, the value of b_0 has little or no practical interpretation.

The slope of price with OmniPower sales ($b_1 = -53.2173$) indicates that, for a given amount of monthly promotional expenditures, the predicted mean sales of OmniPower are estimated to decrease by 53.2173 bars per month for each 1-cent increase in the price. The slope of monthly promotional expenditures with OmniPower sales ($b_2 = 3.6131$) indicates that, for a given price, the predicted mean sales of OmniPower are estimated to increase by 3.6131 bars for each additional \$1 spent on promotions. These estimates allow you to better understand the likely effect that price and promotion decisions will have in the marketplace. For example, a 10-cent decrease in price is predicted to increase mean sales by 532.173 bars, with a fixed amount of monthly promotional expenditures. A \$100 increase in promotional expenditures is predicted to increase mean sales by 361.31 bars for a given price.

Regression coefficients in multiple regression are called **net regression coefficients**, and they estimate the predicted mean change in Y per unit change in a particular X , *holding constant the effect of the other X variables*. For example, in the study of OmniPower bar sales, for a store with a given amount of promotional expenditures, the mean sales are predicted to decrease by 53.2173 bars per month for each 1-cent increase in the price of an OmniPower bar. Another way to interpret this “net effect” is to think of two stores with an equal amount of promotional expenditures. If the first store charges 1 cent more than the other store, the net effect of this difference is that the first store is predicted to sell a mean of 53.2173 fewer bars per month than the second store. To interpret the net effect of promotional expenditures, you can consider two stores that are charging the same price. If the first store spends \$1 more on promotional expenditures, the net effect of this difference is that the first store is predicted to sell a mean of 3.6131 more bars per month than the second store.

Predicting the Dependent Variable Y

You can use the multiple regression equation to predict values of the dependent variable. For example, what are the predicted mean sales for a store charging 79 cents during a month in which promotional expenditures are \$400? Using the multiple regression equation,

$$\hat{Y}_i = 5,837.5208 - 53.2173X_{1i} + 3.6131X_{2i}$$

with $X_{1i} = 79$ and $X_{2i} = 400$,

$$\begin{aligned}\hat{Y}_i &= 5,837.5208 - 53.2173(79) + 3.6131(400) \\ &= 3,078.57\end{aligned}$$

Thus, you predict that stores charging 79 cents and spending \$400 in promotional expenditures will sell a mean of 3,078.57 OmniPower bars per month.

After you have developed the regression equation, done a residual analysis (see Section 14.3), and determined the significance of the overall fitted model (see Section 14.2), you can construct a confidence interval estimate of the mean value and a prediction interval for an individual value. Figure 14.3 presents Excel and Minitab results that compute a confidence interval estimate and a prediction interval for the OmniPower sales data.

Student Tip

Remember that in multiple regression, the regression coefficients are conditional on holding constant the other independent variables. The slope of b_1 holds constant the effect of variable X_2 . The slope of b_2 holds constant the effect of variable X_1 .

Student Tip

You should only predict within the range of the values of all the independent variables.

FIGURE 14.3

Excel confidence interval estimate and prediction interval worksheet for the OmniPower sales data

A	B	C	D
Confidence Interval Estimate and Prediction Interval			
1			
2			
3	Data		
4	Confidence Level	95%	
5		1	
6	Price given value	79	
7	Promotion given value	400	
8			
9	X'X	34 2646 13200	
10		2646 214674 1018800	
11		13200 1018800 6000000	
12			
13	Inverse of X'X	0.9692 -0.0094 -0.0005	
14		-0.0094 0.0001 0.0000	
15		-0.0005 0.0000 0.0000	
16			
17	X'G times Inverse of X'X	0.0121 0.0001 0.0000	
18			
19	[X'G times Inverse of X'X] times XG	0.0298	
20	t Statistic	2.0395	
21	Predicted Y (YHat)	3078.57	
22			
23	For Average Predicted Y (YHat)		
24	Interval Half Width	224.50	
25	Confidence Interval Lower Limit	2854.07	
26	Confidence Interval Upper Limit	3303.08	
27			
28	For Individual Response Y		
29	Interval Half Width	1320.57	
30	Prediction Interval Lower Limit	1758.01	
31	Prediction Interval Upper Limit	4399.14	

Predicted Values for New Observations				
New Obs	Fit	SE Fit	95% CI	95% PI
1	3079	110	(2854, 3303)	(1758, 4399)
Values of Predictors for New Observations				
New Obs	Price	Promotion		
1	79.0	400		

The 95% confidence interval estimate of the mean OmniPower sales for all stores charging 79 cents and spending \$400 in promotional expenditures is 2,854.07 to 3,303.08 bars. The prediction interval for an individual store is 1,758.01 to 4,399.14 bars.

Problems for Section 14.1

LEARNING THE BASICS

14.1 For this problem, use the following multiple regression equation:

$$\hat{Y}_i = 10 + 5X_{1i} + 3X_{2i}$$

- a. Interpret the meaning of the slopes.
- b. Interpret the meaning of the Y intercept.

14.2 For this problem, use the following multiple regression equation:

$$\hat{Y}_i = 50 - 2X_{1i} + 7X_{2i}$$

- a. Interpret the meaning of the slopes.
- b. Interpret the meaning of the Y intercept.

APPLYING THE CONCEPTS

14.3 A shoe manufacturer is considering developing a new brand of running shoes. The business problem facing the marketing analyst is to determine which variables should be used to predict durability (i.e., the effect of long-term impact). Two independent variables under consideration are X_1 (FOREIMP), a measurement of the forefoot shock-absorbing capability, and X_2 (MIDSOLE), a measurement of the change in impact properties over time. The dependent variable Y is LTIMP, a measure of the shoe's durability after a repeated impact test. Data are collected from a random sample of 15 types of currently manufactured running shoes, with the following results:

Variable	Coefficients	Standard Error	t Statistic	p-Value
Intercept	-0.02686	0.06905	-0.39	0.7034
FOREIMP	0.79116	0.06295	12.57	0.0000
MIDSOLE	0.60484	0.07174	8.43	0.0000

- a. State the multiple regression equation.
- b. Interpret the meaning of the slopes, b_1 and b_2 , in this problem.
- c. What conclusions can you reach concerning durability?

14.4 A mail-order catalog business selling personal computer supplies, software, and hardware maintains a centralized warehouse. Management is currently examining the process of distribution from the warehouse. The business problem facing management relates to the factors that affect warehouse distribution costs. Currently, a handling fee is added to each order, regardless of the amount of the order. Data collected over the past 24 months (stored in **WareCost**) indicate the warehouse distribution costs (in \$thousands), the sales (in \$thousands), and the number of orders received.

- a. State the multiple regression equation.
- b. Interpret the meaning of the slopes, b_1 and b_2 , in this problem.
- c. Explain why the regression coefficient, b_0 , has no practical meaning in the context of this problem.
- d. Predict the mean monthly warehouse distribution cost when sales are \$400,000 and the number of orders is 4,500.
- e. Construct a 95% confidence interval estimate for the mean monthly warehouse distribution cost when sales are \$400,000 and the number of orders is 4,500.

- f. Construct a 95% prediction interval for the monthly warehouse distribution cost for a particular month when sales are \$400,000 and the number of orders is 4,500.
- g. Explain why the interval in (e) is narrower than the interval in (f).
- h. What conclusions can you reach concerning warehouse distribution cost?

 **14.5** The production of wine is a multibillion-dollar worldwide industry. In an attempt to develop a model of wine quality as judged by wine experts, data was collected from red wine variants of Portuguese “Vinho Verde” wine. A sample of 50 wines is stored in **VinhoVerde**. (Data extracted from P. Cortez, Cerdeira, A., Almeida, F., Matos, T., and Reis, J., “Modeling Wine Preferences by Data Mining from Physiochemical Properties,” *Decision Support Systems*, 47, 2009, pp. 547–553 and bit.ly/9xKIEa.) Develop a multiple linear regression model to predict wine quality, measured on a scale from 0 (very bad) to 10 (excellent) based on alcohol content (%) and the amount of chlorides.

- a. State the multiple regression equation.
- b. Interpret the meaning of the slopes, b_1 and b_2 , in this problem.
- c. Explain why the regression coefficient, b_0 , has no practical meaning in the context of this problem.
- d. Predict the mean wine quality rating for wines that have 10% alcohol and chlorides of 0.08.
- e. Construct a 95% confidence interval estimate for the mean wine quality rating for wines that have 10% alcohol and chlorides of 0.08.
- f. Construct a 95% prediction interval for the wine quality rating for an individual wine that has 10% alcohol and chlorides of 0.08.
- g. What conclusions can you reach concerning this regression model?

14.6 The business problem facing a consumer products company is to measure the effectiveness of different types of advertising media in the promotion of its products. Specifically, the company is interested in the effectiveness of radio advertising and newspaper advertising (including the cost of discount coupons). During a one-month test period, data were collected from a sample of 22 cities with approximately equal populations. Each city is allocated a specific expenditure level for radio advertising and for newspaper advertising. The sales of the product (in \$thousands) and also the levels of media expenditure (in \$thousands) during the test month are recorded, with the following results shown below and stored in **Advertise**:

City	Sales (\$thousands)	Radio Advertising (\$thousands)	Newspaper Advertising (\$thousands)
16	1,330	55	25
17	1,405	60	30
18	1,436	60	30
19	1,521	65	35
20	1,741	65	35
21	1,866	70	40
22	1,717	70	40

- a. State the multiple regression equation.
- b. Interpret the meaning of the slopes, b_1 and b_2 , in this problem.
- c. Interpret the meaning of the regression coefficient, b_0 .
- d. Which type of advertising is more effective? Explain.

14.7 The business problem facing the director of broadcasting operations for a television station was the issue of standby hours (i.e., hours in which unionized graphic artists at the station are paid but are not actually involved in any activity) and what factors were related to standby hours. The study included the following variables:

Standby hours (Y)—Total number of standby hours in a week
 Total staff present (X_1)—Weekly total of people-days
 Remote hours (X_2)—Total number of hours worked by employees at locations away from the central plant

Data were collected for 26 weeks; these data are organized and stored in **Standby**.

- a. State the multiple regression equation.
- b. Interpret the meaning of the slopes, b_1 and b_2 , in this problem.
- c. Explain why the regression coefficient, b_0 , has no practical meaning in the context of this problem.
- d. Predict the mean standby hours for a week in which the total staff present have 310 people-days and the remote hours total 400.
- e. Construct a 95% confidence interval estimate for the mean standby hours for weeks in which the total staff present have 310 people-days and remote hours total 400.
- f. Construct a 95% prediction interval for the standby hours for a single week in which the total staff present have 310 people-days and the remote hours total 400.
- g. What conclusions can you reach concerning standby hours?

14.8 Nassau County is located approximately 25 miles east of New York City. The data organized and stored in **GlenCove** include the fair market value (in \$thousands), land area of the property in acres, and age, in years, for a sample of 30 single-family homes located in Glen Cove, a small city in Nassau County. Develop a multiple linear regression model to predict the fair market value based on land area of the property and age, in years.

- a. State the multiple regression equation.
- b. Interpret the meaning of the slopes, b_1 and b_2 , in this problem.
- c. Explain why the regression coefficient, b_0 , has no practical meaning in the context of this problem.
- d. Predict the mean fair market value for a house that has a land area of 0.25 acre and is 55 years old.
- e. Construct a 95% confidence interval estimate for the mean fair market value for houses that have a land area of 0.25 acre and are 55 years old.
- f. Construct a 95% prediction interval estimate for the fair market value for an individual house that has a land area of 0.25 acre and is 55 years old.

City	Sales (\$thousands)	Radio Advertising (\$thousands)	Newspaper Advertising (\$thousands)
1	973	0	40
2	1,119	0	40
3	875	25	25
4	625	25	25
5	910	30	30
6	971	30	30
7	931	35	35
8	1,177	35	35
9	882	40	25
10	982	40	25
11	1,628	45	45
12	1,577	45	45
13	1,044	50	0
14	914	50	0
15	1,329	55	25

14.2 r^2 , Adjusted r^2 , and the Overall F Test

This section discusses three methods you can use to evaluate the overall multiple regression model: the coefficient of multiple determination, r^2 , the adjusted r^2 , and the overall F test.

Coefficient of Multiple Determination

Recall from Section 13.3 that the coefficient of determination, r^2 , measures the proportion of the variation in Y that is explained by the independent variable X in the simple linear regression model. In multiple regression, the **coefficient of multiple determination** represents the proportion of the variation in Y that is explained by all the independent variables. Equation (14.4) defines the coefficient of multiple determination for a multiple regression model with two or more independent variables.

COEFFICIENT OF MULTIPLE DETERMINATION

The coefficient of multiple determination is equal to the regression sum of squares (SSR) divided by the total sum of squares (SST).

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST} \quad (14.4)$$

In the OmniPower example, from Figure 14.2 on page 546, $SSR = 39,472,730.77$ and $SST = 52,093,677.44$. Thus,

$$r^2 = \frac{SSR}{SST} = \frac{39,472,730.77}{52,093,677.44} = 0.7577$$

Student Tip

Remember that r^2 in multiple regression represents the proportion of the variation in the dependent variable Y that is explained by all the independent X variables included in the model.

The coefficient of multiple determination ($r^2 = 0.7577$) indicates that 75.77% of the variation in sales is explained by the variation in the price and in the promotional expenditures. The coefficient of multiple determination also appears in the Figure 14.2 results on page 546, labeled R Square in the Excel results and R-Sq in the Minitab results.

Adjusted r^2

When considering multiple regression models, some statisticians suggest that you should use the **adjusted r^2** to take into account both the number of independent variables in the model and the sample size. Reporting the adjusted r^2 is extremely important when you are comparing two or more regression models that predict the same dependent variable but have a different number of independent variables. Equation (14.5) defines the adjusted r^2 .

ADJUSTED r^2

$$r_{\text{adj}}^2 = 1 - \left[(1 - r^2) \frac{n - 1}{n - k - 1} \right] \quad (14.5)$$

where k is the number of independent variables in the regression equation.

Thus, for the OmniPower data, because $r^2 = 0.7577$, $n = 34$, and $k = 2$,

$$\begin{aligned} r_{\text{adj}}^2 &= 1 - \left[(1 - 0.7577) \frac{34 - 1}{34 - 2 - 1} \right] \\ &= 1 - \left[(0.2423) \frac{33}{31} \right] \\ &= 1 - 0.2579 \\ &= 0.7421 \end{aligned}$$

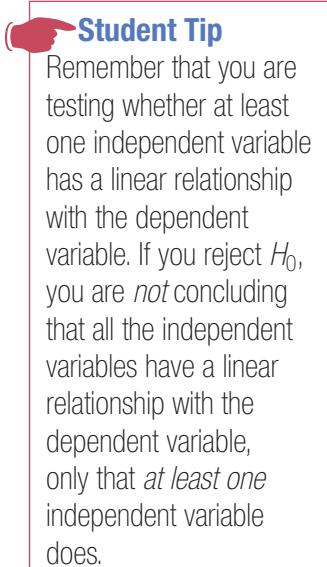
Therefore, 74.21% of the variation in sales is explained by the multiple regression model—adjusted for the number of independent variables and sample size. The adjusted r^2 also appears in the Figure 14.2 results on page 546, labeled Adjusted R Square in the Excel results and R Sq(adj) in the Minitab results.

Test for the Significance of the Overall Multiple Regression Model

You use the **overall F test** to determine whether there is a significant relationship between the dependent variable and the entire set of independent variables (the overall multiple regression model). Because there is more than one independent variable, you use the following null and alternative hypotheses:

$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ (There is no linear relationship between the dependent variable and the independent variables.)

$H_1: \text{At least one } \beta_j \neq 0, j = 1, 2, \dots, k$ (There is a linear relationship between the dependent variable and at least one of the independent variables.)



Equation (14.6) defines the overall F test statistic. Table 14.2 presents the ANOVA summary table.

OVERALL F TEST

The F_{STAT} test statistic is equal to the regression mean square (MSR) divided by the mean square error (MSE).

$$F_{\text{STAT}} = \frac{MSR}{MSE} \quad (14.6)$$

where

k = number of independent variables in the regression model

The F_{STAT} test statistic follows an F distribution with k and $n - k - 1$ degrees of freedom.

TABLE 14.2

ANOVA Summary Table for the Overall F Test

Source	Degrees of Freedom	Sum of Squares	Mean Squares (Variance)	F
Regression	k	SSR	$MSR = \frac{SSR}{k}$	$F_{\text{STAT}} = \frac{MSR}{MSE}$
Error	$n - k - 1$	SSE	$MSE = \frac{SSE}{n - k - 1}$	
Total	$n - 1$	SST		

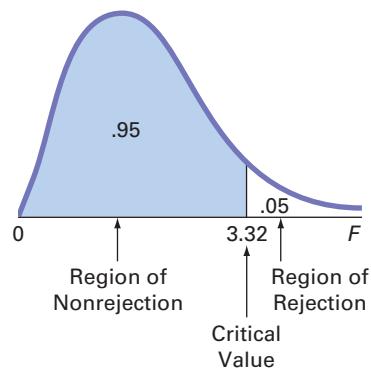
The decision rule is

Reject H_0 at the α level of significance if $F_{STAT} > F_\alpha$;
otherwise, do not reject H_0 .

Using a 0.05 level of significance, the critical value of the F distribution with 2 and 31 degrees of freedom found in Table E.5 is approximately 3.32 (see Figure 14.4 on page 552). From Figure 14.2 on page 546, the F_{STAT} test statistic given in the ANOVA summary table is 48.4771. Because $48.4771 > 3.32$, or because the p -value = 0.000 < 0.05 , you reject H_0 and conclude that at least one of the independent variables (price and/or promotional expenditures) is related to sales.

FIGURE 14.4

Testing for the significance of a set of regression coefficients at the 0.05 level of significance, with 2 and 31 degrees of freedom



Problems for Section 14.2

LEARNING THE BASICS

14.9 The following ANOVA summary table is for a multiple regression model with two independent variables:

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
Regression	2	60		
Error	18	120		
Total	20	180		

- Determine the regression mean square (MSR) and the mean square error (MSE).
- Compute the overall F_{STAT} test statistic.
- Determine whether there is a significant relationship between Y and the two independent variables at the 0.05 level of significance.
- Compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- Compute the adjusted r^2 .

14.10 The following ANOVA summary table is for a multiple regression model with two independent variables:

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
Regression	2	30		
Error	10	120		
Total	12	150		

- Determine the regression mean square (MSR) and the mean square error (MSE).
- Compute the overall F_{STAT} test statistic.
- Determine whether there is a significant relationship between Y and the two independent variables at the 0.05 level of significance.
- Compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- Compute the adjusted r^2 .

APPLYING THE CONCEPTS

14.11 A financial analyst engaged in business valuation obtained financial data on 71 drug companies (Industry Group SIC 3 code: 283). The file **BusinessValuation** contains the following variables:

COMPANY—Drug Company name
PB fye—Price-to-book-value ratio (fiscal year ending)
ROE—Return on equity
SGROWTH—Growth (GS5)

- Develop a regression model to predict price-to-book-value ratio based on return on equity.
- Develop a regression model to predict price-to-book-value ratio based on growth.
- Develop a regression model to predict price-to-book-value ratio based on return on equity and growth.
- Compute and interpret the adjusted r^2 for each of the three models.
- Which of these three models do you think is the best predictor of price-to-book-value ratio?

14.12 In Problem 14.3 on page 548, you predicted the durability of a brand of running shoe, based on the forefoot shock-absorbing capability and the change in impact properties over time. The regression analysis resulted in the following ANOVA summary table:

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F	p-Value
Regression	2	12.61020	6.30510	97.69	0.0001
Error	12	0.77453	0.06454		
Total	14	13.38473			

- a. Determine whether there is a significant relationship between durability and the two independent variables at the 0.05 level of significance.
- b. Interpret the meaning of the *p*-value.
- c. Compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- d. Compute the adjusted r^2 .

14.13 In Problem 14.5 on page 549, you used the percentage of alcohol and chlorides to predict wine quality (stored in **VinhoVerde**). Use the results from that problem to do the following:

- a. Determine whether there is a significant relationship between wine quality and the two independent variables (percentage of alcohol and chlorides) at the 0.05 level of significance.
- b. Interpret the meaning of the *p*-value.
- c. Compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- d. Compute the adjusted r^2 .

 **14.14** In Problem 14.4 on page 548, you used sales and number of orders to predict distribution costs at a mail-order catalog business (stored in **WareCost**). Using the results from that problem,

- a. determine whether there is a significant relationship between distribution costs and the two independent variables (sales and number of orders) at the 0.05 level of significance.

14.3 Residual Analysis for the Multiple Regression Model

In Section 13.5, you used residual analysis to evaluate the fit of the simple linear regression model. For the multiple regression model with two independent variables, you need to construct and analyze the following residual plots:

- Residuals versus \hat{Y}_i
- Residuals versus X_{1i}
- Residuals versus X_{2i}
- Residuals versus time

The first residual plot examines the pattern of residuals versus the predicted values of Y . If the residuals show a pattern for the predicted values of Y , there is evidence of a possible curvilinear effect (see Section 15.1) in at least one independent variable, a possible violation of the assumption of equal variance (see Figure 13.13 on page 510), and/or the need to transform the Y variable.

The second and third residual plots involve the independent variables. Patterns in the plot of the residuals versus an independent variable may indicate the existence of a curvilinear

Student Tip

As is the case with simple linear regression, a residual plot that does not contain any apparent patterns will look like a random scattering of points.

- b. interpret the meaning of the *p*-value.
- c. compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- d. compute the adjusted r^2 .

14.15 In Problem 14.7 on page 549, you used the total staff present and remote hours to predict standby hours (stored in **Standby**). Using the results from that problem,

- a. determine whether there is a significant relationship between standby hours and the two independent variables (total staff present and remote hours) at the 0.05 level of significance.
- b. interpret the meaning of the *p*-value.
- c. compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- d. compute the adjusted r^2 .

14.16 In Problem 14.6 on page 549, you used radio advertising and newspaper advertising to predict sales (stored in **Advertise**). Using the results from that problem,

- a. determine whether there is a significant relationship between sales and the two independent variables (radio advertising and newspaper advertising) at the 0.05 level of significance.
- b. interpret the meaning of the *p*-value.
- c. compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- d. compute the adjusted r^2 .

14.17 In Problem 14.8 on page 549, you used the land area of a property and the age of a house to predict the fair market value (stored in **GlenCove**). Using the results from that problem,

- a. determine whether there is a significant relationship between fair market value and the two independent variables (land area of a property and age of a house) at the 0.05 level of significance.
- b. interpret the meaning of the *p*-value.
- c. compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- d. compute the adjusted r^2 .

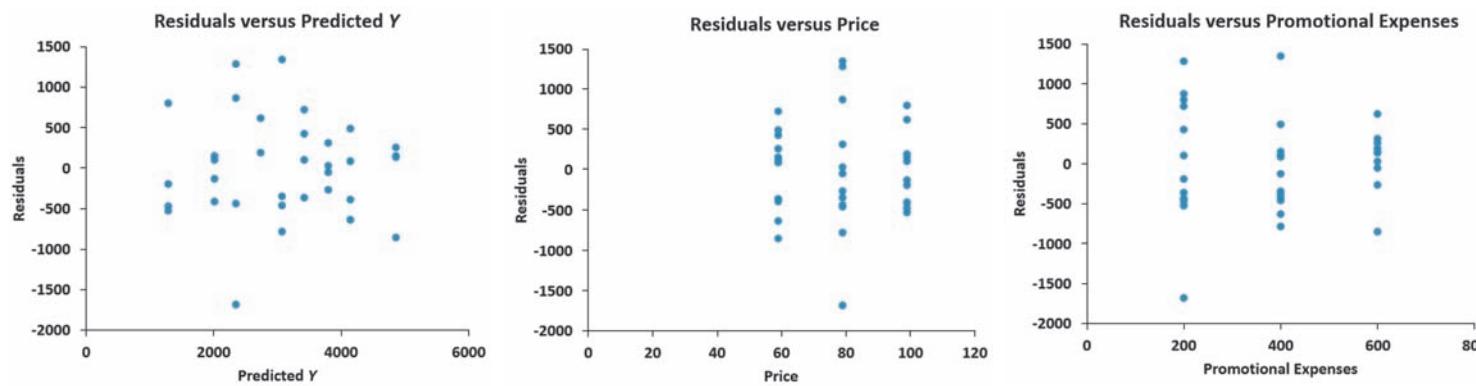
effect and, therefore, the need to add a curvilinear independent variable to the multiple regression model (see Section 15.1).

The fourth plot is used to investigate patterns in the residuals in order to validate the independence assumption when the data are collected in time order. Associated with this residual plot, as in Section 13.6, you can compute the Durbin-Watson statistic to determine the existence of positive autocorrelation among the residuals.

Figure 14.5 presents the residual plots for the OmniPower sales example. There is very little or no pattern in the relationship between the residuals and the predicted value of Y , the value of X_1 (price), or the value of X_2 (promotional expenditures). Thus, you can conclude that the multiple regression model is appropriate for predicting sales. There is no need to plot the residuals versus time because the data were not collected in time order.

FIGURE 14.5

Residual plots for the OmniPower sales data: residuals versus predicted Y , residuals versus price, and residuals versus promotional expenditures



Problems for Section 14.3

APPLYING THE CONCEPTS

14.18 In Problem 14.4 on page 548, you used sales and number of orders to predict distribution costs at a mail-order catalog business (stored in **WareCost**).

- Plot the residuals versus \hat{Y}_i .
- Plot the residuals versus X_{1i} .
- Plot the residuals versus X_{2i} .
- Plot the residuals versus time.
- In the residual plots created in (a) through (d), is there any evidence of a violation of the regression assumptions? Explain.
- Determine the Durbin-Watson statistic.
- At the 0.05 level of significance, is there evidence of positive autocorrelation in the residuals?

14.19 In Problem 14.5 on page 549, you used the percentage of alcohol and chlorides to predict wine quality (stored in **VinhoVerde**).

- Plot the residuals versus \hat{Y}_i .
- Plot the residuals versus X_{1i} .
- Plot the residuals versus X_{2i} .
- In the residual plots created in (a) through (c), is there any evidence of a violation of the regression assumptions? Explain.

e. Should you compute the Durbin-Watson statistic for these data? Explain.

14.20 In Problem 14.6 on page 549, you used radio advertising and newspaper advertising to predict sales (stored in **Advertise**).

- Perform a residual analysis on your results.
- If appropriate, perform the Durbin-Watson test, using $\alpha = 0.05$.
- Are the regression assumptions valid for these data?

14.21 In Problem 14.7 on page 549, you used the total staff present and remote hours to predict standby hours (stored in **Standby**).

- Perform a residual analysis on your results.
- If appropriate, perform the Durbin-Watson test, using $\alpha = 0.05$.
- Are the regression assumptions valid for these data?

14.22 In Problem 14.8 on page 549, you used the land area of a property and the age of a house to predict the fair market value (stored in **GlenCove**).

- Perform a residual analysis on your results.
- If appropriate, perform the Durbin-Watson test, using $\alpha = 0.05$.
- Are the regression assumptions valid for these data?

14.4 Inferences Concerning the Population Regression Coefficients

In Section 13.7, you tested the slope in a simple linear regression model to determine the significance of the relationship between X and Y . In addition, you constructed a confidence interval estimate of the population slope. This section extends those procedures to multiple regression.

Tests of Hypothesis

In a simple linear regression model, to test a hypothesis concerning the population slope, β_1 , you used Equation (13.16) on page 516:

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}}$$

Equation (14.7) generalizes this equation for multiple regression.

TESTING FOR THE SLOPE IN MULTIPLE REGRESSION

$$t_{STAT} = \frac{b_j - \beta_j}{S_{b_j}} \quad (14.7)$$

where

b_j = slope of variable j with Y , holding constant the effects of all other independent variables

S_{b_j} = standard error of the regression coefficient b_j

k = number of independent variables in the regression equation

β_j = hypothesized value of the population slope for variable j , holding constant the effects of all other independent variables

t_{STAT} = test statistic for a t distribution with $n - k - 1$ degrees of freedom

To determine whether variable X_2 (amount of promotional expenditures) has a significant effect on sales, taking into account the price of OmniPower bars, the null and alternative hypotheses are

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

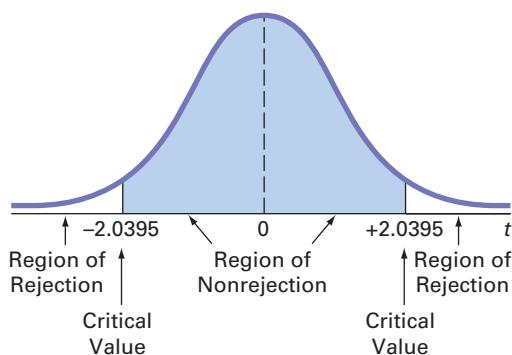
From Equation (14.7) and Figure 14.2 on page 546,

$$\begin{aligned} t_{STAT} &= \frac{b_2 - \beta_2}{S_{b_2}} \\ &= \frac{3.6131 - 0}{0.6852} = 5.2728 \end{aligned}$$

If you select a level of significance of 0.05, the critical values of t for 31 degrees of freedom from Table E.3 are -2.0395 and $+2.0395$ (see Figure 14.6).

FIGURE 14.6

Testing for significance of a regression coefficient at the 0.05 level of significance, with 31 degrees of freedom



From Figure 14.2 on page 546, observe that the computed t_{STAT} test statistic is 5.2728. Because $t_{STAT} = 5.2728 > 2.0395$ or because the p -value is 0.0000, you reject H_0 and conclude that there is a significant relationship between the variable X_2 (promotional expenditures) and sales, taking into account the price, X_1 . The extremely small p -value allows you to strongly reject the null hypothesis that there is no linear relationship between sales and promotional expenditures. Example 14.1 presents the test for the significance of β_1 , the slope of sales with price.

EXAMPLE 14.1

Testing for the Significance of the Slope of Sales with Price

At the 0.05 level of significance, is there evidence that the slope of sales with price is different from zero?

SOLUTION From Figure 14.2 on page 546, $t_{STAT} = -7.7664 < -2.0395$ (the critical value for $\alpha = 0.05$) or the p -value = 0.0000 < 0.05. Thus, there is a significant relationship between price, X_1 , and sales, taking into account the promotional expenditures, X_2 .

As shown with these two independent variables, the test of significance for a specific regression coefficient in multiple regression is a test for the significance of adding that variable into a regression model, given that the other variable is included. In other words, the t test for the regression coefficient is actually a test for the contribution of each independent variable.

Confidence Interval Estimation

Instead of testing the significance of a population slope, you may want to estimate the value of a population slope. Equation (14.8) defines the confidence interval estimate for a population slope in multiple regression.

CONFIDENCE INTERVAL ESTIMATE FOR THE SLOPE

$$b_j \pm t_{\alpha/2} S_{b_j} \quad (14.8)$$

where

$t_{\alpha/2}$ = the critical value corresponding to an upper-tail probability of $\alpha/2$ from the t distribution with $n - k - 1$ degrees of freedom (i.e., a cumulative area of $1 - \alpha/2$)

k = the number of independent variables

To construct a 95% confidence interval estimate of the population slope, β_1 (the effect of price, X_1 , on sales, Y , holding constant the effect of promotional expenditures, X_2), the critical

value of t at the 95% confidence level with 31 degrees of freedom is 2.0395 (see Table E.3). Then, using Equation (14.8) and Figure 14.2 on page 546,

$$\begin{aligned} b_1 &\pm t_{\alpha/2}S_{b_1} \\ -53.2173 &\pm (2.0395)(6.8522) \\ -53.2173 &\pm 13.9752 \\ -67.1925 \leq \beta_1 &\leq -39.2421 \end{aligned}$$

Taking into account the effect of promotional expenditures, the estimated effect of a 1-cent increase in price is to reduce mean sales by approximately 39.2 to 67.2 bars. You have 95% confidence that this interval correctly estimates the relationship between these variables. From a hypothesis-testing viewpoint, because this confidence interval does not include 0, you conclude that the regression coefficient, β_1 , has a significant effect.

Example 14.2 constructs and interprets a confidence interval estimate for the slope of sales with promotional expenditures.

EXAMPLE 14.2

Constructing a Confidence Interval Estimate for the Slope of Sales with Promotional Expenditures

Construct a 95% confidence interval estimate of the population slope of sales with promotional expenditures.

SOLUTION The critical value of t at the 95% confidence level, with 31 degrees of freedom, is 2.0395 (see Table E.3). Using Equation (14.8) and Figure 14.2 on page 546,

$$\begin{aligned} b_2 &\pm t_{\alpha/2}S_{b_2} \\ 3.6131 &\pm (2.0395)(0.6852) \\ 3.6131 &\pm 1.3975 \\ 2.2156 \leq \beta_2 &\leq 5.0106 \end{aligned}$$

Thus, taking into account the effect of price, the estimated effect of each additional dollar of promotional expenditures is to increase mean sales by approximately 2.22 to 5.01 bars. You have 95% confidence that this interval correctly estimates the relationship between these variables. From a hypothesis-testing viewpoint, because this confidence interval does not include 0, you can conclude that the regression coefficient, β_2 , has a significant effect.

Problems for Section 14.4

LEARNING THE BASICS

14.23 Use the following information from a multiple regression analysis:

$$n = 25 \quad b_1 = 5 \quad b_2 = 10 \quad S_{b_1} = 2 \quad S_{b_2} = 8$$

- Which variable has the largest slope, in units of a t statistic?
- Construct a 95% confidence interval estimate of the population slope, β_1 .
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

14.24 Use the following information from a multiple regression analysis:

$$n = 20 \quad b_1 = 4 \quad b_2 = 3 \quad S_{b_1} = 1.2 \quad S_{b_2} = 0.8$$

- Which variable has the largest slope, in units of a t statistic?
- Construct a 95% confidence interval estimate of the population slope, β_1 .
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

APPLYING THE CONCEPTS

14.25 In Problem 14.3 on page 548, you predicted the durability of a brand of running shoe, based on the forefoot shock-absorbing capability (FOREIMP) and the change in impact properties over time (MIDSOLE) for a sample of 15 pairs of shoes. Use the following results:

Variable	Coefficient	Standard Error	t Statistic	p-Value
Intercept	-0.02686	0.06905	-0.39	0.7034
FOREIMP	0.79116	0.06295	12.57	0.0000
MIDSOLE	0.60484	0.07174	8.43	0.0000

- a. Construct a 95% confidence interval estimate of the population slope between durability and forefoot shock-absorbing capability.
- b. At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

 **14.26** In Problem 14.4 on page 548, you used sales and number of orders to predict distribution costs at a mail-order catalog business (stored in **WareCost**). Using the results from that problem,

- a. construct a 95% confidence interval estimate of the population slope between distribution cost and sales.
- b. at the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

14.27 In Problem 14.5 on page 548, you used the percentage of alcohol and chlorides to predict wine quality (stored in **VinhoVerde**). Using the results from that problem,

- a. construct a 95% confidence interval estimate of the population slope between wine quality and the percentage of alcohol.
- b. at the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the

regression model. On the basis of these results, indicate the independent variables to include in this model.

14.28 In Problem 14.6 on page 549, you used radio advertising and newspaper advertising to predict sales (stored in **Advertise**). Using the results from that problem,

- a. construct a 95% confidence interval estimate of the population slope between sales and radio advertising.
- b. at the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

14.29 In Problem 14.7 on page 549, you used the total number of staff present and remote hours to predict standby hours (stored in **Standby**). Using the results from that problem,

- a. construct a 95% confidence interval estimate of the population slope between standby hours and total number of staff present.
- b. at the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

14.30 In Problem 14.8 on page 549, you used land area of a property and age of a house to predict the fair market value (stored in **GlenCove**). Using the results from that problem,

- a. construct a 95% confidence interval estimate of the population slope between fair market value and land area of a property.
- b. at the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

14.5 Testing Portions of the Multiple Regression Model

In developing a multiple regression model, you want to use only those independent variables that significantly reduce the error in predicting the value of a dependent variable. If an independent variable does not improve the prediction, you can delete it from the multiple regression model and use a model with fewer independent variables.

The **partial F test** is an alternative method to the *t* test discussed in Section 14.4 for determining the contribution of an independent variable. Using this method, you determine the contribution to the regression sum of squares made by each independent variable after all the other independent variables have been included in the model. The new independent variable is included only if it significantly improves the model.

To conduct partial *F* tests for the OmniPower sales example, you need to evaluate the contribution of promotional expenditures (X_2) after price (X_1) has been included in the model and also evaluate the contribution of price (X_1) after promotional expenditures (X_2) have been included in the model.

In general, if there are several independent variables, you determine the contribution of each independent variable by taking into account the regression sum of squares of a model that includes all independent variables except the one of interest, j . This regression sum of squares is denoted SSR (all X s except j). Equation (14.9) determines the contribution of variable j , assuming that all other variables are already included.

DETERMINING THE CONTRIBUTION OF AN INDEPENDENT VARIABLE TO THE REGRESSION MODEL

$$SSR(X_j | \text{All } Xs \text{ except } j) = SSR(\text{All } Xs) - SSR(\text{All } Xs \text{ except } j) \quad (14.9)$$

If there are two independent variables, you use Equations (14.10a) and (14.10b) to determine the contribution of each variable.

CONTRIBUTION OF VARIABLE X_1 , GIVEN THAT X_2 HAS BEEN INCLUDED

$$SSR(X_1 | X_2) = SSR(X_1 \text{ and } X_2) - SSR(X_2) \quad (14.10a)$$

CONTRIBUTION OF VARIABLE X_2 , GIVEN THAT X_1 HAS BEEN INCLUDED

$$SSR(X_2 | X_1) = SSR(X_1 \text{ and } X_2) - SSR(X_1) \quad (14.10b)$$

The term $SSR(X_2)$ represents the sum of squares due to regression for a model that includes only the independent variable X_2 (promotional expenditures). Similarly, $SSR(X_1)$ represents the sum of squares due to regression for a model that includes only the independent variable X_1 (price). Figures 14.7 and 14.8 present results for these two models.

FIGURE 14.7

Excel and Minitab results for the simple linear regression model of sales with promotional expenditures, $SSR(X_2)$

A	B	C	D	E	F	G	
1 Sales and Promotional Expenses Analysis							
2							
3 Regression Statistics							
4	Multiple R	0.5351					
5	R Square	0.2863					
6	Adjusted R Square	0.2640					
7	Standard Error	1077.8721					
8	Observations	34					
9							
10 ANOVA							
11	df	SS	MS	F	Significance F		
12	Regression	1	14915814.1025	14915814.1025	12.8384	0.0011	
13	Residual	32	37177863.3387	1161808.2293			
14	Total	33	52093677.4412				
15							
16	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	
17	Intercept	1496.0161	483.9789	3.0911	0.0041	510.1843	2481.8480
18	Promotional Expenses	4.1281	1.1521	3.5831	0.0011	1.7813	6.4748

Regression Analysis: Sales versus Promotion						
The regression equation is						
Sales = 1496 + 4.13 Promotion						
Predictor	Coef	SE Coef	T	P		
Constant	1496.0	484.0	3.09	0.004		
Promotion	4.128	1.152	3.58	0.001		
$S = 1077.87 \quad R-Sq = 28.6\% \quad R-Sq(adj) = 26.4\%$						
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	1	14915814	14915814	12.84	0.001	
Residual Error	32	37177863	1161808			
Total	33	52093677				

FIGURE 14.8

Excel and Minitab results for the simple linear regression model of sales with price, $SSR(X_1)$

A	B	C	D	E	F	G	
1 Sales and Price Analysis							
2							
3 Regression Statistics							
4	Multiple R	0.7351					
5	R Square	0.5404					
6	Adjusted R Square	0.5261					
7	Standard Error	864.9457					
8	Observations	34					
9							
10 ANOVA							
11	df	SS	MS	F	Significance F		
12	Regression	1	28153486.1482	28153486.1482	37.6318	0.0000	
13	Residual	32	23940191.2930	748130.9779			
14	Total	33	52093677.4412				
15							
16	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	
17	Intercept	7512.3480	734.6189	10.2262	0.0000	6015.9796	9008.7164
18	Price	-56.7138	9.2451	-6.1345	0.0000	-75.5455	-37.8822

Regression Analysis: Sales versus Price						
The regression equation is						
Sales = 7512 - 56.7 Price						
Predictor	Coef	SE Coef	T	P		
Constant	7512.3	734.6	10.23	0.000		
Price	-56.714	9.245	-6.13	0.000		
$S = 864.946 \quad R-Sq = 54.0\% \quad R-Sq(adj) = 52.6\%$						
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	1	28153486	28153486	37.63	0.000	
Residual Error	32	23940191	748131			
Total	33	52093677				

From Figure 14.7, $SSR(X_2) = 14,915,814.10$ and from Figure 14.2 on page 546, $SSR(X_1 \text{ and } X_2) = 39,472,730.77$. Then, using Equation (14.10a),

$$\begin{aligned} SSR(X_1 | X_2) &= SSR(X_1 \text{ and } X_2) - SSR(X_2) \\ &= 39,472,730.77 - 14,915,814.10 \\ &= 24,556,916.67 \end{aligned}$$

To determine whether X_1 significantly improves the model after X_2 has been included, you divide the regression sum of squares into two component parts, as shown in Table 14.3.

TABLE 14.3

ANOVA Table Dividing the Regression Sum of Squares into Components to Determine the Contribution of Variable X_1

Source	Degrees of Freedom	Sum of Squares	Mean Square (Variance)	F
Regression	2	39,472,730.77	19,736,365.39	
$\left\{ \begin{array}{l} X_2 \\ X_1 X_2 \end{array} \right\}$	$\left\{ \begin{array}{l} 1 \\ 1 \end{array} \right\}$	$\left\{ \begin{array}{l} 14,915,814.10 \\ 24,556,916.67 \end{array} \right\}$	24,556,916.67	60.32
Error	31	12,620,946.67	407,127.31	
Total	33	52,093,677.44		

The null and alternative hypotheses to test for the contribution of X_1 to the model are:

H_0 : Variable X_1 does not significantly improve the model after variable X_2 has been included.

H_1 : Variable X_1 significantly improves the model after variable X_2 has been included.

Equation (14.11) defines the partial F test statistic for testing the contribution of an independent variable.

PARTIAL F TEST STATISTIC

$$F_{STAT} = \frac{SSR(X_j | \text{All Xs except } j)}{MSE} \quad (14.11)$$

The partial F test statistic follows an F distribution with 1 and $n - k - 1$ degrees of freedom.

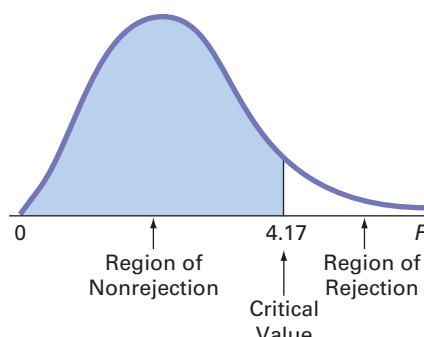
From Table 14.3,

$$F_{STAT} = \frac{24,556,916.67}{407,127.31} = 60.32$$

The partial F_{STAT} test statistic has 1 and $n - k - 1 = 34 - 2 - 1 = 31$ degrees of freedom. Using a level of significance of 0.05, the critical value from Table E.5 is approximately 4.17 (see Figure 14.9).

FIGURE 14.9

Testing for the contribution of a regression coefficient to a multiple regression model at the 0.05 level of significance, with 1 and 31 degrees of freedom



Because the computed partial F_{STAT} test statistic (60.32) is greater than this critical F value (4.17), you reject H_0 . You conclude that the addition of variable X_1 (price) significantly improves a regression model that already contains variable X_2 (promotional expenditures).

To evaluate the contribution of variable X_2 (promotional expenditures) to a model in which variable X_1 (price) has been included, you need to use Equation (14.10b). First, from Figure 14.8 on page 559, observe that $SSR(X_1) = 28,153,486.15$. Second, from Table 14.3, observe that $SSR(X_1 \text{ and } X_2) = 39,472,730.77$. Then, using Equation (14.10b) on page 559,

$$SSR(X_2 | X_1) = 39,472,730.77 - 28,153,486.15 = 11,319,244.62$$

To determine whether X_2 significantly improves a model after X_1 has been included, you can divide the regression sum of squares into two component parts, as shown in Table 14.4.

TABLE 14.4

ANOVA Table Dividing the Regression Sum of Squares into Components to Determine the Contribution of Variable X_2

Source	Degrees of Freedom	Sum of Squares	Mean Square (Variance)	F
Regression	2	39,472,730.77	19,736,365.39	
$\left\{ \begin{array}{l} X_1 \\ X_2 X_1 \end{array} \right\}$	$\left\{ \begin{array}{l} 1 \\ 1 \end{array} \right\}$	$\left\{ \begin{array}{l} 28,153,486.15 \\ 11,319,244.62 \end{array} \right\}$	11,319,244.62	27.80
Error	31	12,620,946.67	407,127.31	
Total	33	52,093,677.44		

The null and alternative hypotheses to test for the contribution of X_2 to the model are:

- H_0 : Variable X_2 does not significantly improve the model after variable X_1 has been included.
 H_1 : Variable X_2 significantly improves the model after variable X_1 has been included.

Using Equation (14.11) and Table 14.4,

$$F_{STAT} = \frac{11,319,244.62}{407,127.31} = 27.80$$

In Figure 14.9, you can see that, using a 0.05 level of significance, the critical value of F , with 1 and 31 degrees of freedom, is approximately 4.17. Because the computed partial F_{STAT} test statistic (27.80) is greater than this critical value (4.17), you reject H_0 . You can conclude that the addition of variable X_2 (promotional expenditures) significantly improves the multiple regression model already containing X_1 (price).

Thus, by testing for the contribution of each independent variable after the other independent variable has been included in the model, you determine that each of the two independent variables significantly improves the model. Therefore, the multiple regression model should include both price, X_1 , and promotional expenditures, X_2 .

The partial F test statistic developed in this section and the t test statistic of Equation (14.7) on page 555 are both used to determine the contribution of an independent variable to a multiple regression model. The hypothesis tests associated with these two statistics always result in the same decision (i.e., the p -values are identical). The t_{STAT} test statistics for the OmniPower regression model are -7.7664 and $+5.2728$, and the corresponding F_{STAT} test statistics are 60.32 and 27.80. Equation (14.12) states this relationship between t and F .¹

¹This relationship holds only when the F_{STAT} statistic has 1 degree of freedom in the numerator.

RELATIONSHIP BETWEEN A t STATISTIC AND AN F STATISTIC

$$t_{STAT}^2 = F_{STAT} \quad (14.12)$$

Student Tip

The coefficients of partial determination measure the proportion of the variation in the dependent variable explained by a specific independent variable, holding the other independent variables constant. They are different from the *coefficient of multiple determination* that measures the proportion of the variation in the dependent variable explained by the entire set of independent variables included in the model.

Coefficients of Partial Determination

Recall from Section 14.2 that the coefficient of multiple determination, r^2 , measures the proportion of the variation in Y that is explained by variation in the independent variables. The **coefficients of partial determination** ($r_{Y1.2}^2$ and $r_{Y2.1}^2$) measure the proportion of the variation in the dependent variable that is explained by each independent variable while controlling for, or holding constant, the other independent variable. Equation (14.13) defines the coefficients of partial determination for a multiple regression model with two independent variables.

COEFFICIENTS OF PARTIAL DETERMINATION FOR A MULTIPLE REGRESSION MODEL CONTAINING TWO INDEPENDENT VARIABLES

$$r_{Y1.2}^2 = \frac{SSR(X_1 | X_2)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_1 | X_2)} \quad (14.13a)$$

and

$$r_{Y2.1}^2 = \frac{SSR(X_2 | X_1)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_2 | X_1)} \quad (14.13b)$$

where

$SSR(X_1 | X_2)$ = sum of squares of the contribution of variable X_1 to the regression model, given that variable X_2 has been included in the model

SST = total sum of squares for Y

$SSR(X_1 \text{ and } X_2)$ = regression sum of squares when variables X_1 and X_2 are both included in the multiple regression model

$SSR(X_2 | X_1)$ = sum of squares of the contribution of variable X_2 to the regression model, given that variable X_1 has been included in the model

For the OmniPower sales example,

$$\begin{aligned} r_{Y1.2}^2 &= \frac{24,556,916.67}{52,093,677.44 - 39,472,730.77 + 24,556,916.67} \\ &= 0.6605 \end{aligned}$$

$$\begin{aligned} r_{Y2.1}^2 &= \frac{11,319,244.62}{52,093,677.44 - 39,472,730.77 + 11,319,244.62} \\ &= 0.4728 \end{aligned}$$

The coefficient of partial determination, $r_{Y1.2}^2$, of variable Y with X_1 while holding X_2 constant is 0.6605. Thus, for a given (constant) amount of promotional expenditures, 66.05% of the variation in OmniPower sales is explained by the variation in the price. The coefficient of partial determination, $r_{Y2.1}^2$, of variable Y with X_2 while holding X_1 constant is 0.4728. Thus, for a given (constant) price, 47.28% of the variation in sales of OmniPower bars is explained by variation in the amount of promotional expenditures.

Equation (14.14) defines the coefficient of partial determination for the j th variable in a multiple regression model containing several (k) independent variables.

COEFFICIENT OF PARTIAL DETERMINATION FOR A MULTIPLE REGRESSION MODEL CONTAINING K INDEPENDENT VARIABLES

$$r_{Y_j,(\text{All variables except } j)}^2 = \frac{SSR(X_j | \text{All Xs except } j)}{SST - SSR(\text{All Xs}) + SSR(X_j | \text{All Xs except } j)} \quad (14.14)$$

Problems for Section 14.5

LEARNING THE BASICS

14.31 The following is the ANOVA summary table for a multiple regression model with two independent variables:

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
Regression	2	60		
Error	18	120		
Total	20	180		

If $SSR(X_1) = 45$ and $SSR(X_2) = 25$,

- determine whether there is a significant relationship between Y and each independent variable at the 0.05 level of significance.
- compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.

14.32 The following is the ANOVA summary table for a multiple regression model with two independent variables:

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
Regression	2	30		
Error	10	120		
Total	12	150		

If $SSR(X_1) = 20$ and $SSR(X_2) = 15$,

- determine whether there is a significant relationship between Y and each independent variable at the 0.05 level of significance.
- compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.

APPLYING THE CONCEPTS

14.33 In Problem 14.5 on page 549, you used alcohol percentage and chlorides to predict wine quality (stored in **VinhoVerde**). Using the results from that problem,

- at the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the most appropriate regression model for this set of data.

- compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.

 **14.34** In Problem 14.4 on page 548, you used sales and number of orders to predict distribution costs at a mail-order catalog business (stored in **WareCost**). Using the results from that problem,

- at the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the most appropriate regression model for this set of data.
- compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.

14.35 In Problem 14.7 on page 549, you used the total staff present and remote hours to predict standby hours (stored in **Standby**). Using the results from that problem,

- at the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the most appropriate regression model for this set of data.
- compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.

14.36 In Problem 14.6 on page 549, you used radio advertising and newspaper advertising to predict sales (stored in **Advertise**). Using the results from that problem,

- at the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the most appropriate regression model for this set of data.
- compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.

14.37 In Problem 14.8 on page 549, you used land area of a property and age of a house to predict the fair market value (stored in **GlenCove**). Using the results from that problem,

- at the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the most appropriate regression model for this set of data.
- compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.

14.6 Using Dummy Variables and Interaction Terms in Regression Models

The multiple regression models discussed in Sections 14.1 through 14.5 assumed that each independent variable is a numerical variable. For example, in Section 14.1, you used price and promotional expenditures, two numerical independent variables, to predict the monthly sales of OmniPower nutrition bars. However, for some models, you need to include the effect of a categorical independent variable. For example, to predict the monthly sales of the OmniPower bars, you might include the categorical variable end-cap location in the model to explore the possible effect on sales caused by displaying the OmniPower bars in the two different end-cap display locations, produce or beverage, used in the North Fork Beverages scenario in Chapter 10.

Dummy Variables

You use a **dummy variable** to include a categorical independent variable in a regression model. A dummy variable X_d recodes the categories of a categorical variable using the numeric values 0 and 1. In the special case of a categorical independent variable that has only two categories, you define one dummy variable, X_d , and use the values 0 and 1 to represent the two categories. For example, for the categorical variable end-cap location discussed in the Chapter 10 Using Statistics scenario, the dummy variable, X_d , would have these values:

$$\begin{aligned} X_d &= 0 \text{ if the observation is in first category (produce end-cap)} \\ X_d &= 1 \text{ if the observation is in second category (beverage end-cap)} \end{aligned}$$

To illustrate using dummy variables in regression, consider the business problem that involves developing a model for predicting the assessed value (\$thousands) of houses in Silver Spring, Maryland, based on house size (in thousands of square feet) and whether the house has a fireplace. To include the categorical variable for the presence of a fireplace, the dummy variable X_2 is defined as

$$\begin{aligned} X_2 &= 0 \text{ if the house does not have a fireplace} \\ X_2 &= 1 \text{ if the house has a fireplace} \end{aligned}$$

Assuming that the slope of assessed value with the size of the house is the same for houses that have and do not have a fireplace, the multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

where

Y_i = assessed value, in thousands of dollars, for house i

β_0 = Y intercept

X_{1i} = house size, in thousands of square feet, for house i

β_1 = slope of assessed value with house size, holding constant the presence or absence of a fireplace

X_{2i} = dummy variable that represents the absence or presence of a fireplace for house i

β_2 = net effect of the presence of a fireplace on assessed value, holding constant the house size

ε_i = random error in Y for house i

Figure 14.10 presents the regression results for this model, using a sample of 30 Silver Spring houses listed for sale that was extracted from [trulia.com](#) and stored in [SilverSpring](#). In these results, the dummy variable X_2 is labeled as FireplaceCoded (Excel) or Fireplace Coded (Minitab).

FIGURE 14.10

Excel and Minitab results for the regression model that includes size of house and presence of fireplace

A	B	C	D	E	F	G
1	Assessed Value Analysis					
2						
Regression Statistics						
4	Multiple R	0.5765				
5	R Square	0.3323				
6	Adjusted R Square	0.2829				
7	Standard Error	61.8788				
8	Observations	30				
9						
10	ANOVA					
11	df	SS	MS	F	Significance F	
12	Regression	2	51462.3960	25731.1980	6.7201	0.0043
13	Residual	27	103382.7120	3828.9893		
14	Total	29	154845.1080			
15						
16	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	269.4185	33.5699	8.0256	0.0000	200.5388
18	Size	49.8215	14.1326	3.5253	0.0015	20.8239
19	FireplaceCoded	12.1623	27.0351	0.4499	0.6564	-43.3092
						67.6339

Regression Analysis: Assessed Value versus Size, Fireplace Coded						
The regression equation is						
Assessed Value = 269 + 49.8 Size + 12.2 Fireplace Coded						
Predictor	Coef	SE Coef	T	P		
Constant	269.42	33.57	8.03	0.000		
Size	49.82	14.13	3.53	0.002		
Fireplace Coded	12.16	27.04	0.45	0.656		
S = 61.8788 R-Sq = 33.2% R-Sq(adj) = 28.3%						
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	2	51462	25731	6.72	0.004	
Residual Error	27	103383	3829			
Total	29	154845				

From Figure 14.10, the regression equation is

$$\hat{Y}_i = 269.4185 + 49.8215X_{1i} + 12.1623X_{2i}$$

For houses without a fireplace, you substitute $X_2 = 0$ into the regression equation:

$$\begin{aligned}\hat{Y}_i &= 269.4185 + 49.8215X_{1i} + 12.1623X_{2i} \\ &= 269.4185 + 49.8215X_{1i} + 12.1623(0) \\ &= 269.4185 + 49.8215X_{1i}\end{aligned}$$

For houses with a fireplace, you substitute $X_2 = 1$ into the regression equation:

$$\begin{aligned}\hat{Y}_i &= 269.4185 + 49.8215X_{1i} + 12.1623X_{2i} \\ &= 269.4185 + 49.8215X_{1i} + 12.1623(1) \\ &= 281.5807 + 49.8215X_{1i}\end{aligned}$$

In this model, the regression coefficients are interpreted as follows:

- Holding constant whether a house has a fireplace, for each increase of 1.0 thousand square feet in house size, the predicted mean assessed value is estimated to increase by 49.8215 thousand dollars (i.e., \$49,821.50).
- Holding constant the house size, the presence of a fireplace is estimated to increase the predicted mean assessed value of the house by 12.1623 thousand dollars (i.e., \$12,162.30).

In Figure 14.10, the t_{STAT} test statistic for the slope of house size with assessed value is 3.5253, and the p -value is 0.015; the t_{STAT} test statistic for presence of a fireplace is 0.4499, and the p -value is 0.6564. Thus, using the 0.05 level of significance, since $0.0015 < 0.05$, the size of the house makes a significant contribution to the model. However, since $0.6564 > 0.05$, the presence of a fireplace does not make a significant contribution to the model. In addition, from Figure 14.10, observe that the coefficient of multiple determination indicates that 33.23% of the variation in assessed value is explained by variation in house size and whether the house has a fireplace. Thus, the variable fireplace does not make a significant contribution and should not be included in the model.

In some situations, the categorical independent variable has more than two categories. When this occurs, two or more dummy variables are needed. Example 14.3 illustrates such a situation.

Student Tip
Remember that an independent variable does not always make a significant contribution to a regression model.

EXAMPLE 14.3

Modeling a Three-Level Categorical Variable

Define a multiple regression model using sales as the dependent variable and package design and price as independent variables. Package design is a three-level categorical variable with designs A , B , or C .

SOLUTION To model the three-level categorical variable package design, two dummy variables, X_1 and X_2 , are needed:

$$\begin{aligned}X_{1i} &= 1 \text{ if package design } A \text{ is used in observation } i; 0 \text{ otherwise} \\ X_{2i} &= 1 \text{ if package design } B \text{ is used in observation } i; 0 \text{ otherwise}\end{aligned}$$

Thus, if observation i uses package design A , then $X_{1i} = 1$ and $X_{2i} = 0$; if observation i uses package design B , then $X_{1i} = 0$ and $X_{2i} = 1$; and if observation i uses package design C , then $X_{1i} = X_{2i} = 0$. Thus, package design C becomes the baseline category to which the effect of package design A and package design B is compared. A third independent variable is used for price:

$$X_{3i} = \text{price for observation } i$$

Thus, the regression model for this example is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

(continued)

where

Y_i = sales for observation i

β_0 = Y intercept

β_1 = difference between the predicted sales of design A and the predicted sales of design C , holding price constant

β_2 = difference between the predicted sales of design B and the predicted sales of design C , holding price constant

β_3 = slope of sales with price, holding the package design constant

ε_i = random error in Y for observation i

Interactions

In the regression models discussed so far, the effect an independent variable has on the dependent variable has been assumed to be independent of the other independent variables in the model. An **interaction** occurs if the effect of an independent variable on the dependent variable changes according to the *value* of a second independent variable. For example, it is possible that advertising will have a large effect on the sales of a product when the price of a product is low. However, if the price of the product is too high, increases in advertising will not dramatically change sales. In this case, price and advertising are said to interact. In other words, you cannot make general statements about the effect of advertising on sales. The effect that advertising has on sales is *dependent* on the price. You use an **interaction term** (sometimes referred to as a **cross-product term**) to model an interaction effect in a regression model.

To illustrate the concept of interaction and use of an interaction term, return to the example concerning the assessed values of homes discussed on pages 564–565. In the regression model, you assumed that the effect that house size has on the assessed value is independent of whether the house has a fireplace. In other words, you assumed that the slope of assessed value with house size is the same for all houses, regardless of whether the house contains a fireplace. If these two slopes are different, an interaction exists between the house size and the presence or absence of a fireplace.

To evaluate whether an interaction exists, you first define an interaction term that is the product of the independent variable X_1 (house size) and the dummy variable X_2 (Fireplace-Coded). You then test whether this interaction variable makes a significant contribution to the regression model. If the interaction is significant, you cannot use the original model for prediction. For these data you define the following:

$$X_3 = X_1 \times X_2$$

Figure 14.11 presents regression results for the model that includes the house size, X_1 , the presence of a fireplace, X_2 , and the interaction of X_1 and X_2 (defined as X_3 and labeled Size*Fireplace).

FIGURE 14.11

Excel and Minitab results for the regression model that includes house size, presence of fireplace, and interaction of house size and fireplace

	A	B	C	D	E	F	G
1	Assessed Value Analysis						
2							
3	Regression Statistics						
4	Multiple R	0.5885					
5	R Square	0.3464					
6	Adjusted R Square	0.2710					
7	Standard Error	62.3909					
8	Observations	30					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	3	53636.8541	17878.9514	4.5930	0.0104	
13	Residual	26	101208.2539	3892.6252			
14	Total	29	154845.1080				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	226.8573	66.2454	3.4245	0.0021	90.6881	363.0266
18	Size	74.7987	36.3298	2.0589	0.0497	0.1216	149.4757
19	FireplaceCoded	63.8028	74.2760	0.8590	0.3982	-88.8737	216.4793
20	Size*Fireplace	-29.5183	39.4946	-0.7474	0.4615	-110.7006	51.6640

Regression Analysis: Assessed Value versus Size, Fireplace Coded, ...						
The regression equation is						
Assessed Value = 227 + 74.8 Size + 63.8 Fireplace Coded						
- 29.5 Size*Fireplace						
Predictor	Coeff	SE Coef	T	P		
Constant	226.86	66.25	3.42	0.002		
Size	74.80	36.33	2.06	0.050		
Fireplace Coded	63.80	74.28	0.86	0.398		
Size*Fireplace	-29.52	39.49	-0.75	0.462		
S = 62.3909	R-Sq = 34.6%	R-Sq(adj) = 27.1%				
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	3	53637	17879	4.59	0.010	
Residual Error	26	101208	3893			
Total	29	154845				

To test for the existence of an interaction, you use the null hypothesis:

$$H_0: \beta_3 = 0$$

versus the alternative hypothesis:

$$H_1: \beta_3 \neq 0.$$

Student Tip

It is possible that the interaction between two independent variables will be significant even though one of the independent variables is not significant.

In Figure 14.11, the t_{STAT} test statistic for the interaction of size and fireplace is -0.7474 . Because $t_{STAT} = -0.7474 > -2.201$ or the p -value $= 0.4615 > 0.05$, you do not reject the null hypothesis. Therefore, the interaction does not make a significant contribution to the model, given that house size and presence of a fireplace are already included. You can conclude that the slope of assessed value with size is the same for houses with fireplaces and houses without fireplaces.

Regression models can have several numerical independent variables along with a dummy variable. Example 14.4 illustrates a regression model in which there are two numerical independent variables and a categorical independent variable.

EXAMPLE 14.4

Studying a Regression Model That Contains a Dummy Variable

The business problem facing a real estate developer involves predicting heating oil consumption in single-family houses. The independent variables considered are atmospheric temperature ($^{\circ}\text{F}$), X_1 , and the amount of attic insulation (inches), X_2 . Data are collected from a sample of 15 single-family houses. Of the 15 houses selected, houses 1, 4, 6, 7, 8, 10, and 12 are ranch-style houses. The data are organized and stored in **HeatingOil**. Develop and analyze an appropriate regression model, using these three independent variables X_1 , X_2 , and X_3 (where X_3 is the dummy variable for ranch-style houses).

SOLUTION Define X_3 , a dummy variable for ranch-style house, as follows:

$$\begin{aligned} X_3 &= 0 \text{ if the style is not ranch} \\ X_3 &= 1 \text{ if the style is ranch} \end{aligned}$$

Assuming that the slope between heating oil consumption and atmospheric temperature, X_1 , and between heating oil consumption and the amount of attic insulation, X_2 , is the same for both styles of houses, the regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

where

Y_i = monthly heating oil consumption, in gallons, for house i

β_0 = Y intercept

β_1 = slope of heating oil consumption with atmospheric temperature, holding constant the effect of attic insulation and the style of the house

β_2 = slope of heating oil consumption with attic insulation, holding constant the effect of atmospheric temperature and the style of the house

β_3 = incremental effect of the presence of a ranch-style house, holding constant the effect of atmospheric temperature and attic insulation

ε_i = random error in Y for house i

Figure 14.12 on page 568 presents results for this regression model.

(continued)

FIGURE 14.12

Excel and Minitab results for the regression model that includes temperature, insulation, and style for the heating oil data

A	B	C	D	E	F	G
1 Heating Oil Consumption Analysis						
2						
3 Regression Statistics						
4	Multiple R	0.9942				
5	R Square	0.9884				
6	Adjusted R Square	0.9853				
7	Standard Error	15.7489				
8	Observations	15				
9						
10	ANOVA					
11	df	SS	MS	F	Significance F	
12	Regression	3	233406.9094	77802.3031	313.6822	0.0000
13	Residual	11	2728.3200	248.0291		
14	Total	14	236135.2293			
15						
16	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	592.5401	14.3370	41.3295	0.0000	560.9846
18	Temperature	-5.5251	0.2044	-27.0267	0.0000	-5.9751
19	Insulation	-21.3761	1.4480	-14.7623	0.0000	-24.5632
20	Ranch-style	-38.9727	8.3584	-4.6627	0.0007	-57.3695

Regression Analysis: Gallons versus Temperature, Insulation, Ranch-style						
The regression equation is						
Gallons = 593 - 5.53 Temperature - 21.4 Insulation - 39.0 Ranch-style						
Predictor	Coef	SE Coef	T	P		
Constant	592.54	14.34	41.33	0.000		
Temperature	-5.5251	0.2044	-27.03	0.000		
Insulation	-21.376	1.448	-14.76	0.000		
Ranch-style	-38.973	8.358	-4.66	0.001		
S = 15.7489	R-Sq = 98.8%	R-Sq(adj) = 98.5%				
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	3	233407	77802	313.68	0.000	
Residual Error	11	2728	248			
Total	14	236135				

From the results in Figure 14.12, the regression equation is

$$\hat{Y}_i = 592.5401 - 5.5251X_{1i} - 21.3761X_{2i} - 38.9727X_{3i}$$

For houses that are not ranch style, because $X_3 = 0$, the regression equation reduces to

$$\hat{Y}_i = 592.5401 - 5.5251X_{1i} - 21.3761X_{2i}$$

For houses that are ranch style, because $X_3 = 1$, the regression equation reduces to

$$\hat{Y}_i = 553.5674 - 5.5251X_{1i} - 21.3761X_{2i}$$

In this model, the regression coefficients are interpreted as follows:

- Holding constant the attic insulation and the house style, for each additional 1°F increase in atmospheric temperature, you estimate that the predicted heating oil consumption decreases by 5.5251 gallons.
- Holding constant the atmospheric temperature and the house style, for each additional 1-inch increase in attic insulation, you estimate that the predicted heating oil consumption decreases by 21.3761 gallons.
- b_3 measures the effect on oil consumption of having a ranch-style house ($X_3 = 1$) compared with having a house that is not ranch style ($X_3 = 0$). Thus, with atmospheric temperature and attic insulation held constant, you estimate that the predicted heating oil consumption is 38.9727 gallons less for a ranch-style house than for a house that is not ranch style.

The three t_{STAT} test statistics representing the slopes for temperature, insulation, and ranch style are -27.0267 , -14.7623 , and -4.6627 . Each of the corresponding p -values is extremely small (less than 0.001). Thus, each of the three variables makes a significant contribution to the model. In addition, the coefficient of multiple determination indicates that 98.84% of the variation in oil usage is explained by variation in temperature, insulation, and whether the house is ranch style.

Before you can use the model in Example 14.4, you need to determine whether the independent variables interact with each other. In Example 14.5, three interaction terms are added to the model.

EXAMPLE 14.5
**Evaluating a
Regression Model
with Several
Interactions**

For the data of Example 14.4, determine whether adding the interaction terms makes a significant contribution to the regression model.

SOLUTION To evaluate possible interactions between the independent variables, three interaction terms are constructed as follows: $X_4 = X_1 \times X_2$, $X_5 = X_1 \times X_3$, and $X_6 = X_2 \times X_3$. The regression model is now

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \varepsilon_i$$

where X_1 is temperature, X_2 is insulation, X_3 is the dummy variable ranch style, X_4 is the interaction between temperature and insulation, X_5 is the interaction between temperature and ranch style, and X_6 is the interaction between insulation and ranch style. Figure 14.13 presents the results for this regression model.

FIGURE 14.13

Excel and Minitab results for the regression model that includes temperature, X_1 ; insulation, X_2 ; the dummy variable ranch-style, X_3 ; the interaction of temperature and insulation, X_4 ; the interaction of temperature and ranch-style, X_5 ; and the interaction of insulation and ranch-style, X_6

A	B	C	D	E	F	G
1 Heating Oil Consumption Analysis						
2						
3 Regression Statistics						
4 Multiple R	0.9966					
5 R Square	0.9931					
6 Adjusted R Square	0.9980					
7 Standard Error	14.2506					
8 Observations	15					
9						
10 ANOVA						
11 df SS MS F Significance F						
12 Regression	6	234510.5818	39085.0970	192.4607	0.0000	
13 Residual	8	1624.6475	203.0809			
14 Total	14	236135.2293				
15						
16 Coefficients Standard Error t Stat P-value Lower 95% Upper 95%						
17 Intercept	642.8867	26.7059	24.0728	0.0000	581.3027	704.4707
18 Temperature	-6.9263	0.7531	-9.1969	0.0000	-8.6629	-5.1896
19 Insulation	-27.8825	3.5801	-7.7882	0.0001	-36.1383	-19.6268
20 Style	-84.6088	29.9956	-2.8207	0.0225	-153.7788	-15.4389
21 Temperature * Insulation	0.1702	0.0886	1.9204	0.0911	-0.0342	0.3746
22 Temperature * Ranch-style	0.6596	0.4617	1.4286	0.1910	-0.4051	1.7242
23 Insulation * Ranch-style	4.9870	3.5137	1.4193	0.1936	-3.1156	13.0895

Regression Analysis: Gallons versus Temperature, Insulation, ...						
The regression equation is						
Gallons = 643 - 6.93 Temperature - 27.9 Insulation						
- 84.6 Ranch-style + 0.170 Temperature*Insulation						
+ 0.660 Temperature*Ranch-style						
+ 4.99 Insulation*Ranch-style						
Predictor Coef SE Coef T P						
Constant 642.89 26.71 24.07 0.000						
Temperature -6.9263 0.7531 -9.20 0.000						
Insulation -27.883 3.580 -7.79 0.000						
Ranch-style -84.61 30.00 -2.82 0.022						
Temperature*Insulation 0.17021 0.08863 1.92 0.091						
Temperature*Ranch-style 0.6596 0.4617 1.43 0.191						
Insulation*Ranch-style 4.987 3.514 1.42 0.194						
S = 14.2506 R-Sq = 99.3% R-Sq(adj) = 98.8%						
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	6	234511	39085	192.46	0.000	
Residual Error	8	1625	203			
Total	14	236135				

To test whether the three interactions significantly improve the regression model, you use the partial F test. The null and alternative hypotheses are

$$H_0: \beta_4 = \beta_5 = \beta_6 = 0 \text{ (There are no interactions among } X_1, X_2, \text{ and } X_3\text{.)}$$

$$H_1: \beta_4 \neq 0 \text{ and/or } \beta_5 \neq 0 \text{ and/or } \beta_6 \neq 0 \text{ (} X_1 \text{ interacts with } X_2, \\ \text{and/or } X_1 \text{ interacts with } X_3, \text{ and/or } X_2 \text{ interacts with } X_3\text{.)}$$

From Figure 14.13,

$$\text{SSR}(X_1, X_2, X_3, X_4, X_5, X_6) = 234,510.5818 \text{ with 6 degrees of freedom}$$

and from Figure 14.12 on page 568, $\text{SSR}(X_1, X_2, X_3) = 233,406.9094$ with 3 degrees of freedom. Thus,

$$\text{SSR}(X_1, X_2, X_3, X_4, X_5, X_6) - \text{SSR}(X_1, X_2, X_3) = 234,510.5818 - 233,406.9094 = 1,103.6724$$

The difference in degrees of freedom is $6 - 3 = 3$.

(continued)

²In general, if a model has several independent variables and you want to test whether additional independent variables contribute to the model, the numerator of the F test is SSR (for all independent variables) minus SSR (for the initial set of variables) divided by the number of independent variables whose contribution is being tested.

To use the partial F test for the simultaneous contribution of three variables to a model, you use an extension of Equation (14.11) on page 560.² The partial F_{STAT} test statistic is

$$F_{STAT} = \frac{[SSR(X_1, X_2, X_3, X_4, X_5, X_6) - SSR(X_1, X_2, X_3)]/3}{MSE(X_1, X_2, X_3, X_4, X_5, X_6)} = \frac{1,103.6724/3}{203.0809} = 1.8115$$

You compare the computed F_{STAT} test statistic to the critical F value for 3 and 8 degrees of freedom. Using a level of significance of 0.05, the critical F value from Table E.5 is 4.07. Because $F_{STAT} = 1.8115 < 4.07$, you conclude that the interactions do not make a significant contribution to the model, given that the model already includes temperature, X_1 ; insulation, X_2 ; and whether the house is ranch style, X_3 . Therefore, the multiple regression model using X_1 , X_2 , and X_3 but no interaction terms is the better model. If you rejected this null hypothesis, you would then test the contribution of each interaction separately in order to determine which interaction terms to include in the model.

Problems for Section 14.6

LEARNING THE BASICS

14.38 Suppose X_1 is a numerical variable and X_2 is a dummy variable with two categories and the regression equation for a sample of $n = 20$ is

$$\hat{Y}_i = 6 + 4X_{1i} + 2X_{2i}$$

- a. Interpret the regression coefficient associated with variable X_1 .
- b. Interpret the regression coefficient associated with variable X_2 .
- c. Suppose that the t_{STAT} test statistic for testing the contribution of variable X_2 is 3.27. At the 0.05 level of significance, is there evidence that variable X_2 makes a significant contribution to the model?

APPLYING THE CONCEPTS

14.39 The chair of the accounting department plans to develop a regression model to predict the grade point average in accounting for those students who are graduating and have completed the accounting major, based on a student's SAT score and whether the student received a grade of B or higher in the introductory statistics course (0 = no and 1 = yes).

- a. Explain the steps involved in developing a regression model for these data. Be sure to indicate the particular models you need to evaluate and compare.
- b. Suppose the regression coefficient for the variable whether the student received a grade of B or higher in the introductory statistics course is +0.30. How do you interpret this result?

14.40 A real estate association in a suburban community would like to study the relationship between the size of a single-family house (as measured by the number of rooms) and the selling price of the house (in \$thousands). Two different neighborhoods are included in the study, one on the east side of the community (=0) and the other on the west side (=1). A random sample of 20 houses was selected, with the results stored in **Neighbor**. For (a) through (k), do not include an interaction term.

- a. State the multiple regression equation that predicts the selling price, based on the number of rooms and the neighborhood.
- b. Interpret the regression coefficients in (a).

- c. Predict the mean selling price for a house with nine rooms that is located in an east-side neighborhood. Construct a 95% confidence interval estimate and a 95% prediction interval.
- d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- e. Is there a significant relationship between selling price and the two independent variables (rooms and neighborhood) at the 0.05 level of significance?
- f. At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- g. Construct and interpret a 95% confidence interval estimate of the population slope for the relationship between selling price and number of rooms.
- h. Construct and interpret a 95% confidence interval estimate of the population slope for the relationship between selling price and neighborhood.
- i. Compute and interpret the adjusted r^2 .
- j. Compute the coefficients of partial determination and interpret their meaning.
- k. What assumption do you need to make about the slope of selling price with number of rooms?
- l. Add an interaction term to the model and, at the 0.05 level of significance, determine whether it makes a significant contribution to the model.
- m. On the basis of the results of (f) and (l), which model is most appropriate? Explain.
- n. What conclusions can the real estate association reach about the effect of the number of rooms and neighborhood on the selling price of homes?

14.41 In Problem 14.5 on page 549, you developed a multiple regression model to predict wine quality for red wines. Now, you wish to determine whether there is an effect on wine quality due to whether the wine is white (0) or red (1). These data are organized and stored in **VinhoVerde-Redandwhite**. Develop a multiple regression model to predict wine quality based on the percentage of alcohol and the type of wine.

For (a) through (m), do not include an interaction term.

- a. State the multiple regression equation that predicts wine quality based on the percentage of alcohol and the type of wine.
 - b. Interpret the regression coefficients in (a).
 - c. Predict the mean quality for a red wine that has 10% alcohol. Construct a 95% confidence interval estimate and a 95% prediction interval.
 - d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.
 - e. Is there a significant relationship between wine quality and the two independent variables (percentage of alcohol and the type of wine) at the 0.05 level of significance?
 - f. At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.
 - g. Construct and interpret 95% confidence interval estimates of the population slope for the relationship between wine quality and the percentage of alcohol and between wine quality and the type of wine.
 - h. Compare the slope in (b) with the slope for the simple linear regression model of Problem 13.4 on page 500. Explain the difference in the results.
 - i. Compute and interpret the meaning of the coefficient of multiple determination, r^2 .
 - j. Compute and interpret the adjusted r^2 .
 - k. Compare r^2 with the r^2 value computed in Problem 13.16 (a) on page 506.
 - l. Compute the coefficients of partial determination and interpret their meaning.
 - m. What assumption about the slope of type of wine with wine quality do you need to make in this problem?
- n. Add an interaction term to the model and, at the 0.05 level of significance, determine whether it makes a significant contribution to the model.
 - o. On the basis of the results of (f) and (n), which model is most appropriate? Explain.
 - p. What conclusions can you reach concerning the effect of alcohol percentage and type of wine on wine quality?

14.42 In mining engineering, holes are often drilled through rock, using drill bits. As a drill hole gets deeper, additional rods are added to the drill bit to enable additional drilling to take place. It is expected that drilling time increases with depth. This increased drilling time could be caused by several factors, including the mass of the drill rods that are strung together. The business problem relates to whether drilling is faster using dry drilling holes or wet drilling holes. Using dry drilling holes involves forcing compressed air down the drill rods to flush the cuttings and drive the hammer. Using wet drilling holes involves forcing water rather than air down the hole. Data have been collected from a sample of 50 drill holes that contains measurements of the time to drill each additional 5 feet (in minutes), the depth (in feet), and whether the hole was a dry drilling hole or a wet drilling hole. The data are organized and stored in **Drill**. (Data extracted from R. Penner and D. G. Watts, "Mining Information," *The American Statistician*, 45, 1991, pp. 4–9.) Develop a model to predict additional drilling time, based on depth and type of drilling hole (dry or wet). For (a) through (k) do not include an interaction term.

- a. State the multiple regression equation.
- b. Interpret the regression coefficients in (a).

- c. Predict the mean additional drilling time for a dry drilling hole at a depth of 100 feet. Construct a 95% confidence interval estimate and a 95% prediction interval.
- d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- e. Is there a significant relationship between additional drilling time and the two independent variables (depth and type of drilling hole) at the 0.05 level of significance?
- f. At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- g. Construct a 95% confidence interval estimate of the population slope for the relationship between additional drilling time and depth.
- h. Construct a 95% confidence interval estimate of the population slope for the relationship between additional drilling time and the type of hole drilled.
- i. Compute and interpret the adjusted r^2 .
- j. Compute the coefficients of partial determination and interpret their meaning.
- k. What assumption do you need to make about the slope of additional drilling time with depth?
- l. Add an interaction term to the model and, at the 0.05 level of significance, determine whether it makes a significant contribution to the model.
- m. On the basis of the results of (f) and (l), which model is most appropriate? Explain.
- n. What conclusions can you reach concerning the effect of depth and type of drilling hole on drilling time?

14.43 The owner of a moving company typically has his most experienced manager predict the total number of labor hours that will be required to complete an upcoming move. This approach has proved useful in the past, but the owner has the business objective of developing a more accurate method of predicting labor hours. In a preliminary effort to provide a more accurate method, the owner has decided to use the number of cubic feet moved and whether there is an elevator in the apartment building as the independent variables and has collected data for 36 moves in which the origin and destination were within the borough of Manhattan in New York City and the travel time was an insignificant portion of the hours worked. The data are organized and stored in **Moving**. For (a) through (k), do not include an interaction term.

- a. State the multiple regression equation for predicting labor hours, using the number of cubic feet moved and whether there is an elevator.
- b. Interpret the regression coefficients in (a).
- c. Predict the mean labor hours for moving 500 cubic feet in an apartment building that has an elevator and construct a 95% confidence interval estimate and a 95% prediction interval.
- d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- e. Is there a significant relationship between labor hours and the two independent variables (cubic feet moved and whether there is an elevator in the apartment building) at the 0.05 level of significance?
- f. At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.

- g. Construct a 95% confidence interval estimate of the population slope for the relationship between labor hours and cubic feet moved.
- h. Construct a 95% confidence interval estimate for the relationship between labor hours and the presence of an elevator.
- i. Compute and interpret the adjusted r^2 .
- j. Compute the coefficients of partial determination and interpret their meaning.
- k. What assumption do you need to make about the slope of labor hours with cubic feet moved?
- l. Add an interaction term to the model, and at the 0.05 level of significance, determine whether it makes a significant contribution to the model.
- m. On the basis of the results of (f) and (l), which model is most appropriate? Explain.
- n. What conclusions can you reach concerning the effect of the number of cubic feet moved and whether there is an elevator on labor hours?



14.44 In Problem 14.4 on page 548, you used sales and orders to predict distribution cost (stored in **WareCost**).

Develop a regression model to predict distribution cost that includes sales, orders, and the interaction of sales and orders.

- a. At the 0.05 level of significance, is there evidence that the interaction term makes a significant contribution to the model?
- b. Which regression model is more appropriate, the one used in (a) or the one used in Problem 14.4? Explain.

14.45 Zagat's publishes restaurant ratings for various locations in the United States. The file **Restaurants** contains the Zagat rating for food, décor, service, and cost per person for a sample of 50 restaurants located in a city and 50 restaurants located in a suburb. (Data extracted from *Zagat Survey 2013, New York City Restaurants*; and *Zagat Survey 2012–2013, Long Island Restaurants*.) Develop a regression model to predict the cost per person, based on a variable that represents the sum of the ratings for food, décor, and service and a dummy variable concerning location (city versus suburban). For (a) through (m), do not include an interaction term.

- a. State the multiple regression equation.
- b. Interpret the regression coefficients in (a).
- c. Predict the mean cost for a restaurant with a summated rating of 60 that is located in a city and construct a 95% confidence interval estimate and a 95% prediction interval.
- d. Perform a residual analysis on the results and determine whether the regression assumptions are satisfied.
- e. Is there a significant relationship between price and the two independent variables (summated rating and location) at the 0.05 level of significance?
- f. At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- g. Construct a 95% confidence interval estimate of the population slope for the relationship between cost and summated rating.
- h. Compare the slope in (b) with the slope for the simple linear regression model of Problem 13.5 on page 500. Explain the difference in the results.
- i. Compute and interpret the meaning of the coefficient of multiple determination.

- j. Compute and interpret the adjusted r^2 .
- k. Compare r^2 with the r^2 value computed in Problem 13.17 (b) on page 506.
- l. Compute the coefficients of partial determination and interpret their meaning.
- m. What assumption about the slope of cost with summated rating do you need to make in this problem?
- n. Add an interaction term to the model and, at the 0.05 level of significance, determine whether it makes a significant contribution to the model.
- o. On the basis of the results of (f) and (n), which model is most appropriate? Explain.
- p. What conclusions can you reach about the effect of the summated rating and the location of the restaurant on the cost of a meal?

14.46 In Problem 14.6 on page 549, you used radio advertising and newspaper advertising to predict sales (stored in **Advertise**). Develop a regression model to predict sales that includes radio advertising, newspaper advertising, and the interaction of radio advertising and newspaper advertising.

- a. At the 0.05 level of significance, is there evidence that the interaction term makes a significant contribution to the model?
- b. Which regression model is more appropriate, the one used in this problem or the one used in Problem 14.6? Explain.

14.47 In Problem 14.5 on page 549, the percentage of alcohol and chlorides were used to predict the quality of red wines (stored in **VinhoVerde**). Develop a regression model that includes the percentage of alcohol, the chlorides, and the interaction of the percentage of alcohol and the chlorides to predict wine quality.

- a. At the 0.05 level of significance, is there evidence that the interaction term makes a significant contribution to the model?
- b. Which regression model is more appropriate, the one used in this problem or the one used in Problem 14.5? Explain.

14.48 In Problem 14.7 on page 549, you used total staff present and remote hours to predict standby hours (stored in **Standby**). Develop a regression model to predict standby hours that includes total staff present, remote hours, and the interaction of total staff present and remote hours.

- a. At the 0.05 level of significance, is there evidence that the interaction term makes a significant contribution to the model?
- b. Which regression model is more appropriate, the one used in this problem or the one used in Problem 14.7? Explain.

14.49 The director of a training program for a large insurance company has the business objective of determining which training method is best for training underwriters. The three methods to be evaluated are classroom, online, and courseware app. The 30 trainees are divided into three randomly assigned groups of 10. Before the start of the training, each trainee is given a proficiency exam that measures mathematics and computer skills. At the end of the training, all students take the same end-of-training exam. The results are organized and stored in **Underwriting**.

Develop a multiple regression model to predict the score on the end-of-training exam, based on the score on the proficiency exam and the method of training used. For (a) through (k), do not include an interaction term.

- a. State the multiple regression equation.
- b. Interpret the regression coefficients in (a).

- c. Predict the mean end-of-training exam score for a student with a proficiency exam score of 100 who had courseware app-based training.
- d. Perform a residual analysis on your results and determine whether the regression assumptions are valid.
- e. Is there a significant relationship between the end-of-training exam score and the independent variables (proficiency score and training method) at the 0.05 level of significance?
- f. At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- g. Construct and interpret a 95% confidence interval estimate of the population slope for the relationship between the end-of-training exam score and the proficiency exam score.
- h. Construct and interpret 95% confidence interval estimates of the population slope for the relationship between the end-of-training exam score and type of training method.
- i. Compute and interpret the adjusted r^2 .
- j. Compute the coefficients of partial determination and interpret their meaning.
- k. What assumption about the slope of proficiency score with end-of-training exam score do you need to make in this problem?
- l. Add interaction terms to the model and, at the 0.05 level of significance, determine whether any interaction terms make a significant contribution to the model.
- m. On the basis of the results of (f) and (l), which model is most appropriate? Explain.

14.7 Logistic Regression

The discussion of the simple linear regression model in Chapter 13 and the multiple regression models in Sections 14.1 through 14.6 only considered *numerical* dependent variables. However, in many applications, the dependent variable is a *categorical* variable that takes on one of only two possible values, such as a customer purchases a product or a customer does not purchase a product. Using a categorical dependent variable violates the normality assumption of the least-squares method and can also result in predicted Y values that are impossible.

An alternative approach to least-squares regression originally applied to survival data in the health sciences (see reference 5), **logistic regression**, enables you to use regression models to predict the probability of a particular categorical response for a given set of independent variables. The logistic regression model uses the **odds ratio**, which represents the probability of an event of interest compared with the probability of not having an event of interest. Equation (14.15) defines the odds ratio.

ODDS RATIO

$$\text{Odds ratio} = \frac{\text{probability of an event of interest}}{1 - \text{probability of an event of interest}} \quad (14.15)$$

Using Equation (14.15), if the probability of an event of interest is 0.50, the odds ratio is

$$\text{Odds ratio} = \frac{0.50}{1 - 0.50} = 1.0, \text{ or } 1 \text{ to } 1$$

If the probability of an event of interest is 0.75, the odds ratio is

$$\text{Odds ratio} = \frac{0.75}{1 - 0.75} = 3.0, \text{ or } 3 \text{ to } 1$$

The logistic regression model is based on the natural logarithm of the odds ratio, $\ln(\text{odds ratio})$.

Equation (14.16) on page 574 defines the logistic regression model for k independent variables.

Student Tip

\ln is the symbol used for natural logarithms, also known as base e logarithms. $\ln(x)$ is the logarithm of x having base e , where $e \approx 2.718282$.

LOGISTIC REGRESSION MODEL

$$\ln(\text{Odds ratio}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (14.16)$$

where

k = number of independent variables in the model

ε_i = random error in observation i

In Sections 13.2 and 14.1, the method of least squares was used to develop a regression equation. In logistic regression, a mathematical method called *maximum likelihood estimation* is typically used to develop a regression equation to predict the natural logarithm of this odds ratio. Equation (14.17) defines the logistic regression equation.

LOGISTIC REGRESSION EQUATION

$$\ln(\text{Estimated odds ratio}) = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki} \quad (14.17)$$

Once you have determined the logistic regression equation, you use Equation (14.18) to compute the estimated odds ratio.

ESTIMATED ODDS RATIO

$$\text{Estimated odds ratio} = e^{\ln(\text{Estimated odds ratio})} \quad (14.18)$$

Once you have computed the estimated odds ratio, you use Equation (14.19) to compute the estimated probability of an event of interest.

ESTIMATED PROBABILITY OF AN EVENT OF INTEREST

$$\text{Estimated probability of an event of interest} = \frac{\text{estimated odds ratio}}{1 + \text{estimated odds ratio}} \quad (14.19)$$

To illustrate the use of logistic regression, consider the case of the sales and marketing manager for the credit card division of a major financial company. The manager wants to conduct a campaign to persuade existing holders of the bank's standard credit card to upgrade, for a nominal annual fee, to the bank's platinum card. The manager wonders, "Which of the existing standard credit cardholders should we target for this campaign?"

The manager has access to the results from a sample of 30 cardholders who were targeted during a pilot campaign last year. These results have been organized as three variables and stored in **CardStudy**. The three variables are cardholder upgraded to a premium card, Y ($0 = \text{no}$, $1 = \text{yes}$); and two independent variables: prior year's credit card purchases (in \$thousands), X_1 ; and cardholder ordered additional credit cards for other authorized users, X_2 ($0 = \text{no}$, $1 = \text{yes}$). Figure 14.14 presents the Excel and Minitab results for the logistic regression model using these data.

FIGURE 14.14

Excel and Minitab logistic regression results for the credit card pilot study data

For other problems, Excel and Minitab results may vary slightly due to the limitations of the Excel methods used in the Section EG14.7 instructions.

A	B	C	D	E
Logistic Regression				
3 Predictor	Coefficients	SE Coef	Z	p-Value
4 Intercept	-6.9394	2.9471	-2.3547	0.0185
5 Purchases	0.1395	0.0681	2.0490	0.0405
6 Extra Cards:1	2.7743	1.1927	2.3261	0.0200
7				
8 Deviance	20.0769			

Binary Logistic Regression: Upgraded versus Purchases, Extra Cards					
Link Function: Logit					
Response Information					
Variable	Value	Count			
Upgraded	1	13	(Event)		
	0	17			
	Total	30			
Logistic Regression Table					
Predictor	Coef	SE Coef	Z	Odds Ratio	95% CI
Constant	-6.93984	2.94712	-2.35	0.019	
Purchases	0.139469	0.0680641	2.05	0.040	1.15 1.01 1.31
Extra Cards	1	2.77434	1.19267	2.33	0.020 16.03 1.55 165.99
Log-Likelihood = -10.038					
Test that all slopes are zero: G = 20.977, DF = 2, P-Value = 0.000					
Goodness-of-Fit Tests					
Method	Chi-Square	DF	P		
Pearson	18.5186	27	0.887		
Deviance	20.0769	27	0.828		
Hosmer-Lemeshow	6.5174	8	0.589		

In this model, the regression coefficients are interpreted as follows:

- The regression constant, b_0 , is -6.9394 . This means that for a credit cardholder who did not charge any purchases last year and who does not have additional cards, the estimated natural logarithm of the odds ratio of purchasing the premium card is -6.9394 .
- The regression coefficient, b_1 , is 0.1395 . This means that holding constant the effect of whether the credit cardholder has additional cards for members of the household, for each increase of \$1,000 in annual credit card spending using the company's card, the estimated natural logarithm of the odds ratio of purchasing the premium card increases by 0.1395 . Therefore, cardholders who charged more in the previous year are more likely to upgrade to a premium card.
- The regression coefficient, b_2 , is 2.7743 . This means that holding constant the annual credit card spending, the estimated natural logarithm of the odds ratio of purchasing the premium card increases by 2.7743 for a credit cardholder who has additional cards for members of the household compared with one who does not have additional cards. Therefore, cardholders possessing additional cards for other members of the household are much more likely to upgrade to a premium card.

The regression coefficients suggest that the credit card company should develop a marketing campaign that targets cardholders who tend to charge large amounts to their cards, and to households that possess more than one card.

As is the case with least-squares regression models, a main purpose of performing logistic regression analysis is to provide predictions of a dependent variable. For example, consider a cardholder who charged \$36,000 last year and possesses additional cards for members of the household. What is the probability the cardholder will upgrade to the premium card during the marketing campaign? Using $X_1 = 36$, $X_2 = 1$, Equation (14.17) on page 574, and the results displayed in Figure 14.14 above,

$$\begin{aligned}\ln(\text{estimated odds of purchasing versus not purchasing}) &= -6.9394 + (0.1395)(36) + (2.7743)(1) \\ &= 0.8569\end{aligned}$$

Then, using Equation (14.18) on page 574,

$$\text{estimated odds ratio} = e^{0.8569} = 2.3558$$

Therefore, the odds are 2.3558 to 1 that a credit cardholder who spent \$36,000 last year and has additional cards will purchase the premium card during the campaign. Using Equation (14.19) on page 574, you can convert this odds ratio to a probability:

$$\begin{aligned}\text{estimated probability of purchasing premium card} &= \frac{2.3558}{1 + 2.3558} \\ &= 0.702\end{aligned}$$

Thus, the estimated probability is 0.702 that a credit cardholder who spent \$36,000 last year and has additional cards will purchase the premium card during the campaign. In other words, you predict 70.2% of such individuals will purchase the premium card.

Now that you have used the logistic regression model for prediction, you need to determine whether or not the model is a good-fitting model. The **deviance statistic** is frequently used to determine whether the current model provides a good fit to the data. This statistic measures the fit of the current model compared with a model that has as many parameters as there are data points (what is called a *saturated* model). The deviance statistic follows a chi-square distribution with $n - k - 1$ degrees of freedom. The null and alternative hypotheses are

Student Tip

Unlike other hypothesis tests, rejecting the null hypothesis here means that the model is *not* a good fit.

H_0 : The model is a good-fitting model.

H_1 : The model is not a good-fitting model.

When using the deviance statistic for logistic regression, the null hypothesis represents a good-fitting model, which is the opposite of the null hypothesis when using the overall F test for the multiple regression model (see Section 14.2). Using the α level of significance, the decision rule is

$$\text{Reject } H_0 \text{ if deviance} > \chi_{\alpha}^2; \\ \text{otherwise, do not reject } H_0.$$

The critical value for a χ^2 statistic with $n - k - 1 = 30 - 2 - 1 = 27$ degrees of freedom is 40.113 (see Table E.4). From Figure 14.14 on page 575, the deviance = 20.0769 < 40.113. Thus, you do not reject H_0 , and you conclude that there is insufficient evidence that the model is not a good-fitting one.

Now that you have concluded that the model is a good-fitting one, you need to evaluate whether each of the independent variables makes a significant contribution to the model in the presence of the others. As is the case with linear regression in Sections 13.7 and 14.4, the test statistic is based on the ratio of the regression coefficient to the standard error of the regression coefficient. In logistic regression, this ratio is defined by the **Wald statistic**, which approximately follows the normal distribution. From Figure 14.14, the Wald statistic (labeled Z) is 2.049 for X_1 and 2.3261 for X_2 . Each of these is greater than the critical value of +1.96 of the normal distribution at the 0.05 level of significance (the p -values are 0.0405 and 0.02). You can conclude that each of the two independent variables makes a contribution to the model in the presence of the other. Therefore, you should include both these independent variables in the model.

Problems for Section 14.7

LEARNING THE BASICS

14.50 Interpret the meaning of a slope coefficient equal to 2.2 in logistic regression.

14.51 Given an estimated odds ratio of 2.5, compute the estimated probability of an event of interest.

14.52 Given an estimated odds ratio of 0.75, compute the estimated probability of an event of interest.

14.53 Consider the following logistic regression equation:

$$\ln(\text{Estimated odds ratio}) = 0.1 + 0.5X_{1i} + 0.2X_{2i}$$

- Interpret the meaning of the logistic regression coefficients.
- If $X_1 = 2$ and $X_2 = 1.5$, compute the estimated odds ratio and interpret its meaning.
- On the basis of the results of (b), compute the estimated probability of an event of interest.

APPLYING THE CONCEPTS

14.54 Refer to Figure 14.14 on page 575.

- Predict the probability that a cardholder who charged \$36,000 last year and does not have any additional credit cards for members of the household will purchase the platinum card during the marketing campaign.
- Compare the results in (a) with those for a person with additional credit cards.
- Predict the probability that a cardholder who charged \$18,000 and does not have any additional credit cards for other authorized users will purchase the platinum card during the marketing campaign.
- Compare the results of (a) and (c) and indicate what implications these results might have for the strategy for the marketing campaign.

14.55 A study was conducted to determine the factors involved in the rate of participation of discharged cardiac patients in a rehabilitation program. Data were collected from 516 treated patients (data extracted from F. Van Der Meulen, T. Vermaat, and P. Williams, “Case Study: An Application of Logistic Regression in a Six Sigma Project in Health Care,” *Quality Engineering*, 2011, pp. 113–124). Among the variables used to predict participation (0 = no, 1 = yes) were the distance traveled to rehabilitation in kilometers, whether the person had a car (0 = no, 1 = yes), and the age of the person in years. The summarized data are:

	Estimate	Standard Error	Z Value	<i>p</i> -value
Intercept	5.7765	0.8619	6.702	0.0000
Distance	−0.0675	0.0111	−6.113	0.0000
Car	1.9369	0.2720	7.121	0.0000
Age	−0.0599	0.0119	−5.037	0.0000

- a. State the logistic regression model.
- b. Using the model in (a), predict the probability that a patient will participate in rehabilitation if he or she travels 20 km to rehabilitation, has a car, and is 65 years old.
- c. Using the model in (a), predict the probability that a patient will participate in rehabilitation if he or she travels 20 km to rehabilitation, does not have a car, and is 65 years old.
- d. Compare the results of (b) and (c).
- e. At the 0.05 level of significance, is there evidence that the distance traveled, whether the patient has a car, and the age of the patient each make a significant contribution to the model?
- f. What conclusions can you reach about the likelihood of a patient participating in the rehabilitation program?

14.56 Referring to Problem 14.41 on page 570, you have decided to analyze whether there are differences in fixed acidity, chlorides, and pH between white wines and red wines (0 = white 1 = red). Using the data stored in **RedandWhite**,

- a. Develop a logistic regression model to predict whether the wine is red based on the fixed acidity, chlorides, and pH.
- b. Explain the meaning of the regression coefficients in the model developed in (a).
- c. Predict the probability that a wine is red if it has a fixed acidity of 7.0, chlorides of 0.04, and pH of 3.5.
- d. At the 0.05 level of significance, is there evidence that the logistic regression model developed in (a) is a good fitting model?
- e. At the 0.05 level of significance, is there evidence that fixed acidity, chlorides, and pH each make a significant contribution to the model?
- f. What conclusions concerning the probability of a wine selected being red can you reach?

14.57 Undergraduate students at Miami University in Oxford, Ohio, were surveyed in order to evaluate the effect of price on the purchase of a pizza from Pizza Hut. The students were asked to suppose that they were going to have a large two-topping pizza delivered to their residence. Then they were asked to select from either Pizza Hut or another pizzeria of their choice. The price they would have to pay to get a Pizza Hut pizza differed from survey to survey. For example, some surveys used the price \$11.49. Other prices investigated were \$8.49, \$9.49, \$10.49, \$12.49, \$13.49, and

\$14.49. The dependent variable for this study is whether or not a student will select Pizza Hut. Possible independent variables are the price of a Pizza Hut pizza and the gender of the student. The file **PizzaHut** contains responses from 220 students and includes these three variables:

Gender: 1 = male, 0 = female

Price: 8.49, 9.49, 10.49, 11.49, 12.49, 13.49, or 14.49

Purchase: 1 = the student selected Pizza Hut, 0 = the student selected another pizzeria

- a. Develop a logistic regression model to predict the probability that a student selects Pizza Hut based on the price of the pizza. Is price an important indicator of purchase selection?
- b. Develop a logistic regression model to predict the probability that a student selects Pizza Hut based on the price of the pizza and the gender of the student. Is price an important indicator of purchase selection? Is gender an important indicator of purchase selection?
- c. Compare the results from (a) and (b). Which model would you choose? Discuss.
- d. Using the model selected in (c), predict the probability that a student will select Pizza Hut if the price is \$8.99.
- e. Using the model selected in (c), predict the probability that a student will select Pizza Hut if the price is \$11.49.
- f. Using the model selected in (c), predict the probability that a student will select Pizza Hut if the price is \$13.99.

14.58 An automotive insurance company wants to predict which filed stolen vehicle claims are fraudulent, based on the mean number of claims submitted per year by the policy holder and whether the policy is a new policy, that is, is one year old or less (coded as 1 = yes, 0 = no). Data from a random sample of 98 automotive insurance claims, organized and stored in **InsuranceFraud**, show that 49 are fraudulent (coded as 1) and 49 are not (coded as 0). (Data extracted from A. Gepp *et al.*, “A Comparative Analysis of Decision Trees vis-à-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection,” *Journal of Data Science*, 10 (2012), pp. 537–561.)

- a. Develop a logistic regression model to predict the probability of a fraudulent claim, based on the number of claims submitted per year by the policy holder and whether the policy is new.
- b. Explain the meaning of the regression coefficients in the model in (a).
- c. Predict the probability of a fraudulent claim given that the policy holder has submitted a mean of one claim per year and holds a new policy.
- d. At the 0.05 level of significance, is there evidence that a logistic regression model that uses the mean number of claims submitted per year by the policy holder and whether the policy is new to predict the probability of a fraudulent claim is a good fitting model?
- e. At the 0.05 level of significance, is there evidence that the mean number of claims submitted per year by the policy holder and whether the policy is new each makes a significant contribution to the logistic model?
- f. Develop a logistic regression model that includes only the number of claims submitted per year by the policy holder to predict the probability of a fraudulent claim.
- g. Develop a logistic regression model that includes only whether the policy is new to predict a fraudulent claim.
- h. Compare the models in (a), (f), and (g). Evaluate the differences among the models.

14.59 A marketing manager wants to predict customers with the risk of churning (switching their service contracts to another company) based on the number of calls the customer makes to the company call center and the number of visits the customer makes to the local service center. Data from a random sample of 30 customers, organized and stored in **Churn**, show that 15 have churned (coded as 1) and 15 have not (coded as 0)

- Develop a logistic regression model to predict the probability of churn, based on the number of calls the customer makes to the company call center and the number of visits the customer makes to the local service center.
- Explain the meaning of the regression coefficients in the model in (a).
- Predict the probability of churn for a customer who called the company call center 10 times and visited the local service center once.

- At the 0.05 level of significance, is there evidence that a logistic regression model that uses the number of calls the customer makes to the company call center and the number of visits the customer makes to the local service center is a good fitting model?
- At the 0.05 level of significance, is there evidence that the number of calls the customer makes to the company call center and the number of visits the customer makes to the local service center each make a significant contribution to the logistic model?
- Develop a logistic regression model that includes only the number of calls the customer makes to the company call center to predict the probability of churn.
- Develop a logistic regression model that includes only the number of visits the customer makes to the local service center to predict churn.
- Compare the models in (a), (f), and (g). Evaluate the differences among the models.

14.8 Influence Analysis

In Sections 13.5 and 14.3, you used residual analysis to evaluate the regression assumptions. This section introduces several methods that measure the influence of individual values:

- The hat matrix elements, h_i
- The Studentized deleted residuals, t_i
- Cook's distance statistic, D_i

Figure 14.15 presents these statistics for the OmniPower sales data, as computed by Minitab.

FIGURE 14.15

Minitab worksheet containing computed values for the Studentized deleted residuals (labeled TRES1), the hat matrix elements, and Cook's distance statistics for the OmniPower sales data

#	C1	C2	C3	C4	C5	C6
	Sales	Price	Promotional Expenses	TRES1	H11	COOK1
1	4141	59		200	1.21248	0.119048
2	3842	59		200	0.69829	0.119048
3	3056	59		200	-0.60203	0.119048
4	3519	59		200	0.16218	0.119048
5	4226	59		400	0.13285	0.069940
6	4630	59		400	0.78667	0.069940
7	3507	59		400	-1.03461	0.069940
8	3754	59		400	-0.62580	0.069940
9	5000	59		600	0.22031	0.113095
10	5120	59		600	0.41780	0.113095
11	4011	59		600	-1.44695	0.113095
12	5015	59		600	0.24494	0.113095
13	1916	79		200	-0.70923	0.069940
14	675	79		200	-3.08402	0.069940
15	3636	79		200	2.20612	0.069940
16	3224	79		200	1.43451	0.069940
17	2295	79		400	-1.25842	0.029762
18	2730	79		400	-0.54832	0.029762
19	2618	79		400	-0.72723	0.029762
20	4421	79		400	2.27527	0.029762
21	4113	79		600	0.50383	0.081845
22	3746	79		600	-0.08881	0.081845
23	3532	79		600	-0.43448	0.081845
24	3825	79		600	0.03832	0.081845
25	1096	99		200	-0.32079	0.113095
26	761	99		200	-0.87981	0.113095
27	2088	99		200	1.34235	0.113095
28	820	99		200	-0.77987	0.113095
29	2114	99		400	0.16060	0.081845
30	1882	99		400	-0.21292	0.081845
31	2159	99		400	0.23315	0.081845
32	1602	99		400	-0.66819	0.081845
33	3354	99		600	1.04633	0.142857
34	2927	99		600	0.31720	0.142857

The Hat Matrix Elements, h_i

In Section 13.8, h_i was defined for the simple linear regression model when constructing the confidence interval estimate of the mean response. For multiple regression models, the equation for calculating the **hat matrix diagonal elements**, h_i , requires the use of matrix algebra and is beyond the scope of this text (see references 1, 2, and 6).

The hat matrix diagonal element for observation i , denoted h_i , reflects the possible influence of X_i on the regression equation. If potentially influential observations are present, you may need to delete them from the model. In a regression model containing k independent variables, Hoaglin and Welsch (see reference 6) suggest the following decision rule:

$$\text{If } h_i > 2(k + 1)/n,$$

then X_i is an influential observation and is a candidate for removal from the model.

For the OmniPower sales data, because $n = 34$ and $k = 2$, you flag any h_i value greater than $2(2 + 1)/34 = 0.1765$. Referring to Figure 14.15, you see that none of the h_i values are greater than 0.1429. Therefore, none of the observations are candidates for removal from the analysis.

The Studentized Deleted Residuals, t_i

Recall from Section 13.5 that a *residual* is the difference between the observed value of Y and the predicted value of Y [see Equation (13.14) on page 507]. Studentized residuals are the residuals divided by the standard error of the estimate S_{YX} and adjusted for the distance from \bar{X} . The **Studentized deleted residual**, expressed as a *t* statistic in Equation (14.20), measures the difference of each Y_i , from the value predicted by a model that includes all observations *except* observation i .

STUDENTIZED DELETED RESIDUAL

$$t_i = e_i \sqrt{\frac{n - k - 1}{SSE(1 - h_i) - e_i^2}} \quad (14.20)$$

where

e_i = residual for observation i

k = number of independent variables

SSE = error sum of squares of the regression model fitted

h_i = hat matrix diagonal element for observation i

Hoaglin and Welsch (see reference 6) suggest that if $t_i > t_{\alpha/2}$ or $t_i < -t_{\alpha/2}$ (using a level of significance of 0.10), the observed and predicted values are so different that observation i is highly influential on the regression equation and is a candidate for removal.

For the OmniPower sales data, $n = 34$ and $k = 2$. Thus, you flag any t_i whose absolute value is greater than 1.6973 (see Table E.3). In Figure 14.15, $t_{14} = -3.08402$, $t_{15} = 2.20612$, and $t_{20} = 2.27527$ are highlighted. Thus, the 14th, 15th, and 20th observations may each have an adverse effect on the model. These observations were not previously flagged according to the h_i criterion. Since h_i and t_i measure different aspects of influence, neither criterion is sufficient by itself. When h_i is small, t_i may be large. When h_i is large, t_i may be moderate or small because the observed Y_i is consistent with the rest of the data.

Cook's Distance Statistic, D_i

Cook's distance statistic, D_i , based on both h_i and the Studentized residual, is a third criterion for identifying influential observations. To decide whether an observation flagged by either the h_i or t_i criterion is unduly affecting the model, Cook and Weisberg (see reference 4) developed Cook's D_i statistic.

COOK'S D_i STATISTIC

$$D_i = \frac{e_i^2}{k \text{MSE}} \left[\frac{h_i}{(1 - h_i)^2} \right] \quad (14.21)$$

where

 e_i = residual for observation i k = number of independent variables MSE = mean square error of the regression model fitted h_i = hat matrix diagonal element for observation i

Cook and Weisberg suggest that if $D_i > F_\alpha$ (the critical value of the F distribution having $k + 1$ degrees of freedom in the numerator and $n - k - 1$ degrees of freedom in the denominator at a 0.50 level of significance), the observation is highly influential on the regression equation and is a candidate for removal.

Table 14.5 shows critical values for Cook's D_i statistic.

TABLE 14.5Selected Critical Values of F for Cook's D_i Statistic

$\alpha = 0.50$													
<i>Numerator df = k + 1</i>													
Denominator	$df = n - k - 1$	2	3	4	5	6	7	8	9	10	12	15	20
10	.743	.845	.899	.932	.954	.971	.983	.992	1.00	1.01	1.02	1.03	
11	.739	.840	.893	.926	.948	.964	.977	.986	.994	1.01	1.02	1.03	
12	.735	.835	.888	.921	.943	.959	.972	.981	.989	1.00	1.01	1.02	
15	.726	.826	.878	.911	.933	.949	.960	.970	.977	.989	1.00	1.01	
20	.718	.816	.868	.900	.922	.938	.950	.959	.966	.977	.989	1.00	
24	.714	.812	.863	.895	.917	.932	.944	.953	.961	.972	.983	.994	
30	.709	.807	.858	.890	.912	.927	.939	.948	.955	.966	.978	.989	
40	.705	.802	.854	.885	.907	.922	.934	.943	.950	.961	.972	.983	
60	.701	.798	.849	.880	.901	.917	.928	.937	.945	.956	.967	.978	
120	.697	.793	.844	.875	.896	.912	.923	.932	.939	.950	.961	.972	
∞	.693	.789	.839	.870	.891	.907	.918	.927	.934	.945	.956	.967	

For the OmniPower sales data, since $n = 34$ and $k = 2$, there are 3 degrees of freedom in the numerator and 31 degrees of freedom in the denominator. Thus, any $D_i > 0.807$ is flagged. Referring to Figure 14.15, you see that none of the D_i values exceed 0.187, and therefore no observations are identified as influential using Cook's D_i statistic.

Comparison of Statistics

For the OmniPower sales data, the three statistics do not lead to a consistent set of conclusions about the influence of each observation on the multiple regression model. According to both the h_i and the D_i criteria, none of the observations is a candidate for removal. Under such circumstances, most statisticians would conclude that there is insufficient evidence for the removal of any observation from the analysis.

Individual statisticians may show a preference for a particular statistic to evaluate the influence of each observation on the multiple regression model, including statistics not discussed in this section (see references 1 and 3). Currently, no consensus exists as to which statistic is the best one to use.

Problems for Section 14.8

APPLYING THE CONCEPTS

14.60 In Problem 14.4 on page 548, you used sales and number of orders to predict distribution costs at a mail-order catalog business (stored in **Warecost**). Perform an influence analysis on your results and determine whether any observations should be deleted from the analysis. If necessary, reanalyze the regression model after deleting these observations and compare your results.

14.61 In Problem 14.5 on page 549, you used the percentage of alcohol and the chlorides to predict wine quality (stored in **VinhoVerde**). Perform an influence analysis on your results and determine whether any observations should be deleted from the analysis. If necessary, reanalyze the regression model after deleting these observations and compare your results.

14.62 In Problem 14.6 on page 549, you used the amount of radio advertising and newspaper advertising to predict sales (stored

in **Advertise**). Perform an influence analysis on your results and determine whether any observations should be deleted from the analysis. If necessary, reanalyze the regression model after deleting these observations and compare your results.

14.63 In Problem 14.7 on page 549, you used the total staff present and remote hours to predict standby hours (stored in **Standby**). Perform an influence analysis on your results and determine whether any observations should be deleted from the analysis. If necessary, reanalyze the regression model after deleting these observations and compare your results.

14.64 In Problem 14.8 on page 549, you used the land area of the property and age in years to predict fair market value (stored in **GlenCove**). Perform an influence analysis on your results and determine whether any observations should be deleted from the analysis. If necessary, reanalyze the regression model after deleting these observations and compare your results.

USING STATISTICS

The Multiple Effects of OmniPower Bars, Revisited

In the Using Statistics scenario, you were a marketing manager for OmniFoods, responsible for nutrition bars and similar snack items. You needed to determine the effect that price and in-store promotions would have on sales of OmniPower nutrition bars in order to develop an effective marketing strategy. A sample of 34 stores in a supermarket chain was selected for a test-market study. The stores charged between 59 and 99 cents per bar and were given an in-store promotion budget between \$200 and \$600.

At the end of the one-month test-market study, you performed a multiple regression analysis on the data. Two independent variables were considered: the price of an OmniPower bar and the monthly budget for in-store promotional expenditures. The dependent variable was the number of OmniPower bars sold in a month. The coefficient of determination indicated that 75.8% of the variation in sales was explained by knowing the price charged and the amount spent on in-store promotions. The model indicated that the predicted sales of OmniPower are estimated to decrease by 532

bars per month for each 10-cent increase in the price, and the predicted sales are estimated to increase by 361 bars for each additional \$100 spent on promotions.

After studying the relative effects of price and promotion, OmniFoods needs to set price and promotion standards for a nationwide introduction (obviously, lower prices and higher promotion budgets lead to more sales, but they do so at a lower profit margin). You determined that if stores spend \$400 a month for in-store promotions and charge 79 cents, the 95% confidence interval estimate of the mean monthly sales is 2,854 to 3,303 bars. OmniFoods can multiply the lower and upper bounds of this confidence interval by the number of stores included in the nationwide introduction to estimate total monthly sales. For example, if 1,000 stores are in the nationwide introduction, then total monthly sales should be between 2.854 million and 3.308 million bars.



Ariwasabi/Shutterstock

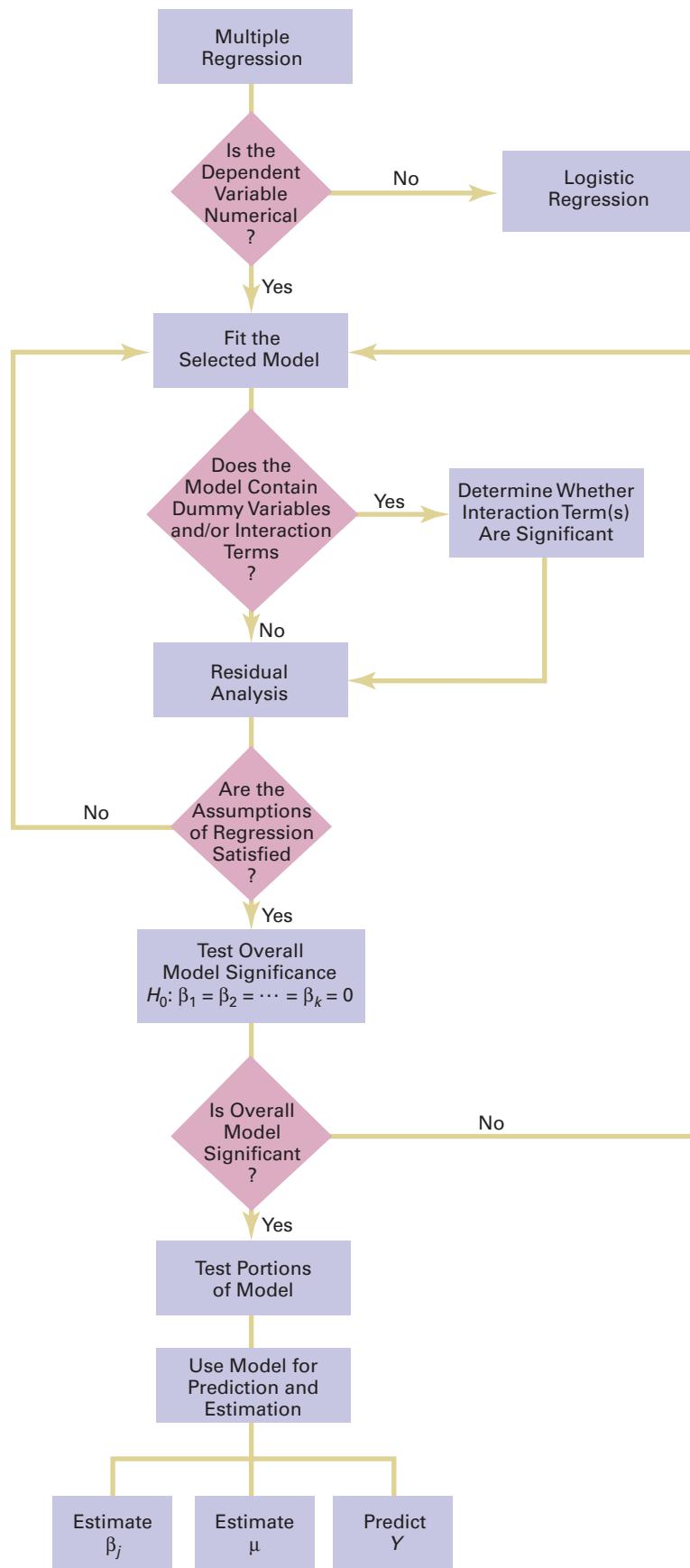
SUMMARY

Figure 14.16 presents a roadmap of this chapter. In this chapter, you learned how to develop and fit multiple regression models that use two or more independent variables to predict the value of a dependent variable. You also

learned how to include categorical independent variables and interaction terms in regression models and learned the logistic regression model that is used to predict a categorical dependent variable.

FIGURE 14.16

Roadmap for multiple regression



REFERENCES

1. Andrews, D. F., and D. Pregibon. "Finding the Outliers that Matter." *Journal of the Royal Statistical Society* 40 (Ser. B., 1978): 85–93.
2. Atkinson, A. C. "Robust and Diagnostic Regression Analysis." *Communications in Statistics* 11 (1982): 2559–2572.
3. Belsley, D. A., E. Kuh, and R. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley, 1980.
4. Cook, R. D., and S. Weisberg. *Residuals and Influence in Regression*. New York: Chapman and Hall, 1982.
5. Hosmer, D. W., and S. Lemeshow. *Applied Logistic Regression*, 2nd ed. New York: Wiley, 2001.
6. Hoaglin, D. C., and R. Welsch. "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, (1978), 17–22.
7. Kutner, M., C. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*, 5th ed. New York: McGraw-Hill/Irwin, 2005.
8. Microsoft Excel 2013. Redmond, WA: Microsoft Corp., 2012.
9. Minitab Release 16. State College, PA: Minitab, Inc., 2010.

KEY EQUATIONS

Multiple Regression Model with k Independent Variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (14.1)$$

Multiple Regression Model with Two Independent Variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (14.2)$$

Multiple Regression Equation with Two Independent Variables

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} \quad (14.3)$$

Coefficient of Multiple Determination

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST} \quad (14.4)$$

Adjusted r^2

$$r_{\text{adj}}^2 = 1 - \left[(1 - r^2) \frac{n - 1}{n - k - 1} \right] \quad (14.5)$$

Overall F Test

$$F_{\text{STAT}} = \frac{MSR}{MSE} \quad (14.6)$$

Testing for the Slope in Multiple Regression

$$t_{\text{STAT}} = \frac{b_j - \beta_j}{S_{b_j}} \quad (14.7)$$

Confidence Interval Estimate for the Slope

$$b_j \pm t_{\alpha/2} S_{b_j} \quad (14.8)$$

Determining the Contribution of an Independent Variable to the Regression Model

$$SSR(X_j | \text{All Xs except } j) = SSR(\text{All Xs}) - SSR(\text{All Xs except } j) \quad (14.9)$$

Contribution of Variable X_1 , Given That X_2 Has Been Included

$$SSR(X_1 | X_2) = SSR(X_1 \text{ and } X_2) - SSR(X_2) \quad (14.10a)$$

Contribution of Variable X_2 , Given That X_1 Has Been Included

$$SSR(X_2 | X_1) = SSR(X_1 \text{ and } X_2) - SSR(X_1) \quad (14.10b)$$

Partial F Test Statistic

$$F_{\text{STAT}} = \frac{SSR(X_j | \text{All Xs except } j)}{MSE} \quad (14.11)$$

Relationship Between a t Statistic and an F Statistic

$$t_{\text{STAT}}^2 = F_{\text{STAT}} \quad (14.12)$$

Coefficients of Partial Determination for a Multiple Regression Model Containing Two Independent Variables

$$r_{Y1,2}^2 = \frac{SSR(X_1 | X_2)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_1 | X_2)} \quad (14.13a)$$

and

$$r_{Y2,1}^2 = \frac{SSR(X_2 | X_1)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_2 | X_1)} \quad (14.13b)$$

Coefficient of Partial Determination for a Multiple Regression Model Containing k Independent Variables

$$r_{Y_j(\text{All variables except } j)}^2 = \frac{SSR(X_j | \text{All Xs except } j)}{SST - SSR(\text{All Xs}) + SSR(X_j | \text{All Xs except } j)} \quad (14.14)$$

Odds Ratio

$$\text{Odds ratio} = \frac{\text{probability of an event of interest}}{1 - \text{probability of an event of interest}} \quad (14.15)$$

Logistic Regression Model

$$\ln(\text{Odds ratio}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (14.16)$$

Logistic Regression Equation

$$\ln(\text{Estimated odds ratio}) = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki} \quad (14.17)$$

Estimated Odds Ratio

$$\text{Estimated odds ratio} = e^{\ln(\text{Estimated odds ratio})} \quad (14.18)$$

Estimated Probability of an Event of Interest

Estimated probability of an event of interest

$$= \frac{\text{estimated odds ratio}}{1 + \text{estimated odds ratio}} \quad (14.19)$$

Studentized Deleted Residual

$$t_i = e_i \sqrt{\frac{n - k - 1}{SSE(1 - h_i) - e_i^2}} \quad (14.20)$$

Cook's D_i Statistic

$$D_i = \frac{e_i^2}{k \text{MSE}} \left[\frac{h_i}{(1 - h_i)^2} \right] \quad (14.21)$$

KEY TERMS

adjusted r^2 550
 coefficient of multiple determination 549
 coefficient of partial determination 562
 Cook's distance statistic D_i 579
 cross-product term 566
 deviance statistic 576

dummy variable 564
 hat matrix diagonal elements h_i 579
 interaction 566
 interaction term 566
 logistic regression 573
 multiple regression model 544

net regression coefficient 547
 odds ratio 573
 overall F test 551
 partial F test 558
 Studentized deleted residual 579
 Wald statistic 576

CHECKING YOUR UNDERSTANDING

14.65 What is the difference between r^2 and adjusted r^2 ?

14.66 How does the interpretation of the regression coefficients differ in multiple regression and simple linear regression?

14.67 How does testing the significance of the entire multiple regression model differ from testing the contribution of each independent variable?

14.68 How do the coefficients of partial determination differ from the coefficient of multiple determination?

14.69 Why and how do you use dummy variables?

14.70 How can you evaluate whether the slope of the dependent

variable with an independent variable is the same for each level of the dummy variable?

14.71 Under what circumstances do you include an interaction term in a regression model?

14.72 When a dummy variable is included in a regression model that has one numerical independent variable, what assumption do you need to make concerning the slope between the dependent variable, Y , and the numerical independent variable, X ?

14.73 When do you use logistic regression?

14.74 What is the difference between the hat matrix diagonal elements h_i and the Studentized deleted residuals?

CHAPTER REVIEW PROBLEMS

14.75 Increasing customer satisfaction typically results in increased purchase behavior. For many products, there is more than one measure of customer satisfaction. In many, purchase behavior can increase dramatically with an increase in just one of the customer satisfaction measures. Gunst and Barry ("One Way to Moderate Ceiling Effects," *Quality Progress*, October 2003, pp. 83–85) consider a product with two satisfaction measures, X_1 and X_2 , that range from the lowest level of satisfaction, 1, to the highest level of satisfaction, 7. The dependent variable, Y , is a measure of purchase behavior, with the highest value generating the most sales. Consider the regression equation:

$$\hat{Y}_i = -3.888 + 1.449X_{1i} + 1.462X_{2i} - 0.190X_{1i}X_{2i}$$

Suppose that X_1 is the perceived quality of the product and X_2 is the perceived value of the product. (Note: If the customer thinks the product is overpriced, he or she perceives it to be of low value and vice versa.)

- What is the predicted purchase behavior when $X_1 = 2$ and $X_2 = 2$?
- What is the predicted purchase behavior when $X_1 = 2$ and $X_2 = 7$?
- What is the predicted purchase behavior when $X_1 = 7$ and $X_2 = 2$?
- What is the predicted purchase behavior when $X_1 = 7$ and $X_2 = 7$?
- What is the regression equation when $X_2 = 2$? What is the slope for X_1 now?
- What is the regression equation when $X_2 = 7$? What is the slope for X_1 now?
- What is the regression equation when $X_1 = 2$? What is the slope for X_2 now?
- What is the regression equation when $X_1 = 7$? What is the slope for X_2 now?
- Discuss the implications of (a) through (h) in the context of increasing sales for this product with two customer satisfaction measures.

14.76 The owner of a moving company typically has his most experienced manager predict the total number of labor hours that will be required to complete an upcoming move. This approach has proved useful in the past, but the owner has the business objective of developing a more accurate method of predicting labor hours. In a preliminary effort to provide a more accurate method, the owner has decided to use the number of cubic feet moved and the number of pieces of large furniture as the independent variables and has collected data for 36 moves in which the origin and destination were within the borough of Manhattan in New York City and the travel time was an insignificant portion of the hours worked. The data are organized and stored in **Moving**.

- State the multiple regression equation.
- Interpret the meaning of the slopes in this equation.
- Predict the mean labor hours for moving 500 cubic feet with two large pieces of furniture.
- Perform a residual analysis on your results and determine whether the regression assumptions are valid.
- Determine whether there is a significant relationship between labor hours and the two independent variables (the number of cubic feet moved and the number of pieces of large furniture) at the 0.05 level of significance.
- Determine the p -value in (e) and interpret its meaning.
- Interpret the meaning of the coefficient of multiple determination in this problem.
- Determine the adjusted r^2 .
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- Determine the p -values in (i) and interpret their meaning.
- Construct a 95% confidence interval estimate of the population slope between labor hours and the number of cubic feet moved. How does the interpretation of the slope here differ from that in Problem 13.44 on page 521?
- Compute and interpret the coefficients of partial determination.
- What conclusions can you reach concerning labor hours?

14.77 Professional basketball has truly become a sport that generates interest among fans around the world. More and more players come from outside the United States to play in the National Basketball Association (NBA). You want to develop a regression model to predict the number of wins achieved by each NBA team, based on field goal (shots made) percentage for the team and for the opponent. The data are stored in **NBA2012**.

- State the multiple regression equation.
- Interpret the meaning of the slopes in this equation.
- Predict the mean number of wins for a team that has a field goal percentage of 45% and an opponent field goal percentage of 44%.
- Perform a residual analysis on your results and determine whether the regression assumptions are valid.
- Is there a significant relationship between number of wins and the two independent variables (field goal percentage for the team and for the opponent) at the 0.05 level of significance?
- Determine the p -value in (e) and interpret its meaning.
- Interpret the meaning of the coefficient of multiple determination in this problem.
- Determine the adjusted r^2 .
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the

regression model. Indicate the most appropriate regression model for this set of data.

- Determine the p -values in (i) and interpret their meaning.
- Compute and interpret the coefficients of partial determination.
- Perform an influence analysis on your results and determine whether any observations should be deleted from the model. If necessary, reanalyze the regression model after deleting these observations and compare your results to those of the original model.
- What conclusions can you reach concerning field goal percentage (team and opponent) in predicting the number of wins?

14.78 A sample of 30 houses recently listed for sale in Silver Spring, Maryland, was selected with the objective of developing a model to predict the assessed value (in \$thousands), using the size of the house (in thousands of square feet) and age (in years). The results are stored in **Silver Spring**.

- Fit a multiple regression model.
- Interpret the meaning of the slopes in this model.
- Predict the mean assessed value for a house that has 2,000 square feet and is 55 years old.
- Perform a residual analysis on your results and determine whether the regression assumptions are valid.
- Determine whether there is a significant relationship between assessed value and the two independent variables (house size and age) at the 0.05 level of significance.
- Determine the p -value in (e) and interpret its meaning.
- Interpret the meaning of the coefficient of multiple determination in this problem.
- Determine the adjusted r^2 .
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- Determine the p -values in (i) and interpret their meaning.
- Construct a 95% confidence interval estimate of the population slope between assessed value and the size of the house. How does the interpretation of the slope here differ from that in Problem 13.76 on page 533?
- Compute and interpret the coefficients of partial determination.
- What conclusions can you reach about the assessed value?

14.79 Measuring the height of a California redwood tree is very difficult because these trees grow to heights over 300 feet. People familiar with these trees understand that the height of a California redwood tree is related to other characteristics of the tree, including the diameter of the tree at the breast height of a person (in inches) and the thickness of the bark of the tree (in inches). The file **Redwood** contains the height, diameter at breast height of a person, and bark thickness for a sample of 21 California redwood trees.

- State the multiple regression equation that predicts the height of a tree, based on the tree's diameter at breast height and the thickness of the bark.
- Interpret the meaning of the slopes in this equation.
- Predict the mean height for a tree that has a breast height diameter of 25 inches and a bark thickness of 2 inches.
- Interpret the meaning of the coefficient of multiple determination in this problem.
- Perform a residual analysis on the results and determine whether the regression assumptions are valid.

- f. Determine whether there is a significant relationship between the height of redwood trees and the two independent variables (breast-height diameter and bark thickness) at the 0.05 level of significance.
- g. Construct a 95% confidence interval estimate of the population slope between the height of redwood trees and breast-height diameter and between the height of redwood trees and the bark thickness.
- h. At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. Indicate the independent variables to include in this model.
- i. Construct a 95% confidence interval estimate of the mean height for trees that have a breast-height diameter of 25 inches and a bark thickness of 2 inches, along with a prediction interval for an individual tree.
- j. Compute and interpret the coefficients of partial determination.
- k. Perform an influence analysis on your results and determine whether any observations should be deleted from the model. If necessary, reanalyze the regression model after deleting these observations and compare your results to those of the original model.
- l. What conclusions can you reach concerning the effect of the diameter of the tree and the thickness of the bark on the height of the tree?

14.80 A sample of 30 houses recently listed for sale in Silver Spring, Maryland, was selected with the objective of developing a model to predict the taxes (in \$) based on the assessed value of houses (in \$thousands) and the age of the houses (in years) (stored in **SilverSpring**):

- a. State the multiple regression equation.
- b. Interpret the meaning of the slopes in this equation.
- c. Predict the mean taxes for a house that has an assessed value of \$400,000 and is 50 years old.
- d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- e. Determine whether there is a significant relationship between taxes and the two independent variables (assessed value and age) at the 0.05 level of significance.
- f. Determine the p -value in (e) and interpret its meaning.
- g. Interpret the meaning of the coefficient of multiple determination in this problem.
- h. Determine the adjusted r^2 .
- i. At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- j. Determine the p -values in (i) and interpret their meaning.
- k. Construct a 95% confidence interval estimate of the population slope between taxes and assessed value. How does the interpretation of the slope here differ from that of Problem 13.77 on page 533?
- l. Compute and interpret the coefficients of partial determination.
- m. The real estate assessor's office has been publicly quoted as saying that the age of a house has no bearing on its taxes. Based on your answers to (a) through (l), do you agree with this statement? Explain.

14.81 A baseball analytics specialist wants to determine which variables are important in predicting a team's wins in a given season. He has collected data related to wins, earned run average

(ERA), and runs scored for the 2012 season (stored in **BB2012**). Develop a model to predict the number of wins based on ERA and runs scored.

- a. State the multiple regression equation.
- b. Interpret the meaning of the slopes in this equation.
- c. Predict the mean number of wins for a team that has an ERA of 4.50 and has scored 750 runs.
- d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- e. Is there a significant relationship between the number of wins and the two independent variables (ERA and runs scored) at the 0.05 level of significance?
- f. Determine the p -value in (e) and interpret its meaning.
- g. Interpret the meaning of the coefficient of multiple determination in this problem.
- h. Determine the adjusted r^2 .
- i. At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- j. Determine the p -values in (i) and interpret their meaning.
- k. Construct a 95% confidence interval estimate of the population slope between wins and ERA.
- l. Compute and interpret the coefficients of partial determination.
- m. Which is more important in predicting wins—pitching, as measured by ERA, or offense, as measured by runs scored? Explain.
- n. Perform an influence analysis on your results and determine whether any observations should be deleted from the model. If necessary, reanalyze the regression model after deleting these observations and compare your results to those of the original model.

14.82 Referring to Problem 14.81, suppose that in addition to using ERA to predict the number of wins, the analytics specialist wants to include the league (0 = American, 1 = National) as an independent variable. Develop a model to predict wins based on ERA and league. For (a) through (k), do not include an interaction term.

- a. State the multiple regression equation.
- b. Interpret the slopes in (a).
- c. Predict the mean number of wins for a team with an ERA of 4.50 in the American League. Construct a 95% confidence interval estimate for all teams and a 95% prediction interval for an individual team.
- d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- e. Is there a significant relationship between wins and the two independent variables (ERA and league) at the 0.05 level of significance?
- f. At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- g. Construct a 95% confidence interval estimate of the population slope for the relationship between wins and ERA.
- h. Construct a 95% confidence interval estimate of the population slope for the relationship between wins and league.
- i. Compute and interpret the adjusted r^2 .
- j. Compute and interpret the coefficients of partial determination.
- k. What assumption do you have to make about the slope of wins with ERA?

- I. Add an interaction term to the model and, at the 0.05 level of significance, determine whether it makes a significant contribution to the model.
- m. On the basis of the results of (f) and (l), which model is most appropriate? Explain.

14.83 You are a real estate broker who wants to compare property values in Glen Cove and Roslyn (which are located approximately 8 miles apart). In order to do so, you will analyze the data in **GCRoslyn**, a file that includes samples of houses from Glen Cove and Roslyn. Making sure to include the dummy variable for location (Glen Cove or Roslyn), develop a regression model to predict fair market value, based on the land area of a property, the age of a house, and location. Be sure to determine whether any interaction terms need to be included in the model.

14.84 A recent article discussed a metal deposition process in which a piece of metal is placed in an acid bath and an alloy is layered on top of it. The business objective of engineers working on the process was to reduce variation in the thickness of the alloy layer. To begin, the temperature and the pressure in the tank holding the acid bath are to be studied as independent variables. Data are collected from 50 samples. The results are organized and stored in **Thickness**. (Data extracted from J. Conklin, "It's a Marathon, Not a Sprint," *Quality Progress*, June 2009, pp. 46–49.)

Develop a multiple regression model that uses temperature and the pressure in the tank holding the acid bath to predict the thickness of the alloy layer. Be sure to perform a thorough residual analysis. The article suggests that there is a significant interaction between the pressure and the temperature in the tank. Do you agree?

14.85 Starbucks Coffee Co. uses a data-based approach to improving the quality and customer satisfaction of its products. When survey data indicated that Starbucks needed to improve its package

sealing process, an experiment was conducted to determine the factors in the bag-sealing equipment that might be affecting the ease of opening the bag without tearing the inner liner of the bag. (Data extracted from L. Johnson and S. Burrows, "For Starbucks, It's in the Bag," *Quality Progress*, March 2011, pp. 17–23.) Among the factors that could affect the rating of the ability of the bag to resist tears were the viscosity, pressure, and plate gap on the bag-sealing equipment. Data were collected on 19 bags in which the plate gap was varied. The results are stored in **Starbucks**. Develop a multiple regression model that uses the viscosity, pressure, and plate gap on the bag-sealing equipment to predict the tear rating of the bag. Be sure to perform a thorough residual analysis. Do you think that you need to use all three independent variables in the model? Explain.

14.86 An experiment was conducted to study the extrusion process of biodegradable packaging foam (data extracted from W. Y. Koh, K. M. Eskridge, and M. A. Hanna, "Supersaturated Split-Plot Designs," *Journal of Quality Technology*, 45, January 2013, pp. 61–72). Among the factors considered for their effect on the unit density (mg/ml) were the die temperature (145°C versus 155°C) and the die diameter (3 mm versus 4 mm. The results were stored in **PackagingFoam3**. Develop a multiple regression model that uses die temperature and die diameter to predict the unit density (mg/ml). Be sure to perform a thorough residual and influence analysis. Do you think that you need to use both independent variables in the model? Explain.

14.87 Referring to Problem 14.86, instead of predicting the unit density, you now wish to predict the foam diameter from results stored in **PackagingFoam4**. Develop a multiple regression model that uses die temperature and die diameter to predict the foam diameter (mg/ml). Be sure to perform a thorough residual and influence analysis. Do you think that you need to use both independent variables in the model? Explain.

CASES FOR CHAPTER 14

Managing Ashland MultiComm Services

In its continuing study of the 3-For-All subscription solicitation process, a marketing department team wants to test the effects of two types of structured sales presentations (personal formal and personal informal) and the number of hours spent on telemarketing on the number of new subscriptions. The staff has recorded these data for the past 24 weeks in **AMS14**.

Analyze these data and develop a multiple regression model to predict the number of new subscriptions for a week, based on the number of hours spent on telemarketing and the sales presentation type. Write a report, giving detailed findings concerning the regression model used.

Digital Case

Apply your knowledge of multiple regression models in this Digital Case, which extends the OmniFoods Using Statistics scenario from this chapter.

To ensure a successful test marketing of its OmniPower energy bars, the OmniFoods marketing department has contracted with In-Store Placements Group (ISPG), a merchandising consulting firm. ISPG will work with the grocery store chain that is conducting the test-market study. Using the same 34-store sample used in the test-market study, ISPG claims that the choice of shelf location and the presence of in-store OmniPower coupon dispensers both increase sales of the energy bars.

Open **Omni_ISPGMemo.pdf** to review the ISPG claims and supporting data. Then answer the following questions:

1. Are the supporting data consistent with ISPG's claims? Perform an appropriate statistical analysis to confirm (or discredit) the stated relationship between sales and the two independent variables of product shelf location and the presence of in-store OmniPower coupon dispensers.
2. If you were advising OmniFoods, would you recommend using a specific shelf location and in-store coupon dispensers to sell OmniPower bars?
3. What additional data would you advise collecting in order to determine the effectiveness of the sales promotion techniques used by ISPG?

CHAPTER 14 EXCEL GUIDE

EG14.1 DEVELOPING a MULTIPLE REGRESSION MODEL

Interpreting the Regression Coefficients

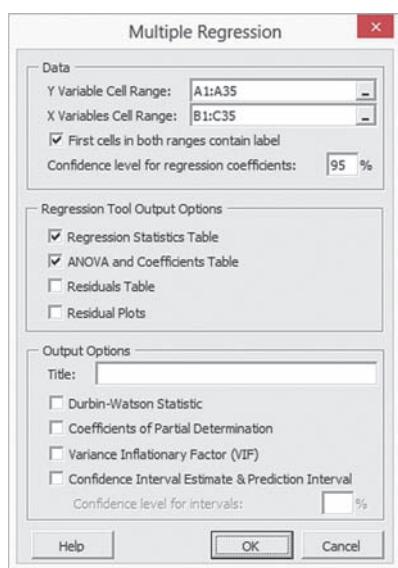
Key Technique Use the **LINEST(cell range of Y variable, cell range of X variables, True, True)** function to compute the regression coefficients and other values related to a multiple regression analysis.

Example Develop the Figure 14.2 multiple regression model for the OmniPower sales data shown on page 546.

PHStat Use Multiple Regression.

For the example, open to the **DATA worksheet** of the **OmniPower workbook**. Select **PHStat → Regression → Multiple Regression**, and in the procedure's dialog box (shown below):

1. Enter **A1:A35** as the **Y Variable Cell Range**.
2. Enter **B1:C35** as the **X Variables Cell Range**.
3. Check **First cells in both ranges contain label**.
4. Enter **95** as the **Confidence level for regression coefficients**.
5. Check **Regression Statistics Table** and **ANOVA and Coefficients Table**.
6. Enter a **Title** and click **OK**.



The procedure creates a worksheet that contains a copy of your data in addition to the Figure 14.2 worksheet. For more information about these worksheets, read the following *In-Depth Excel* section.

In-Depth Excel Use the **COMPUTE worksheet** of the **Multiple Regression workbook** as a template.

For the example, the COMPUTE worksheet uses the OmniPower sales data already in the **MRData worksheet** to perform the regression analysis. To perform multiple regression analyses for other data, paste the regression data into the MRData worksheet.

Figure 14.2 does not show the Calculations area in columns K through N. In the cell range L2:N6, an array formula uses the **LINEST** function to compute intercepts, standard error values, and other regression statistics. The Calculations area also contains the user-supplied confidence level and formulas to compute the critical value of the *t* statistic and half-widths.

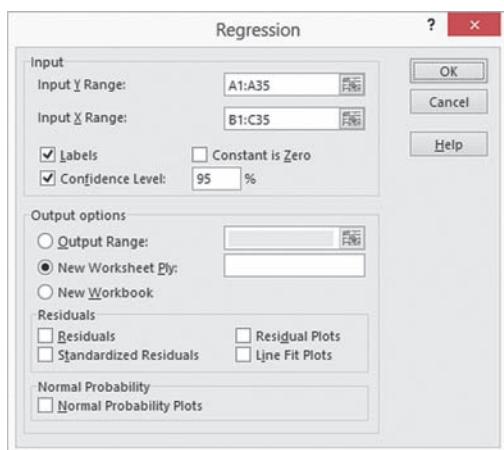
To perform a multiple regression analysis with other data, paste the regression data into the MRData worksheet. Paste the values for the *Y* variable into column A and the values for the *X* variables into consecutive columns, starting with column B. Then, open to the COMPUTE worksheet. Enter the confidence level in cell L8 and edit the five-row-by-three-column array formula that starts with cell L2 (the cell range L2:N6). If you have more than two independent variables, select the wider range that adds a column for each independent variable in excess of two. For example, with three independent variables, select the cell range L2:O6. Then, edit the array formula to reflect the data you pasted into the MRData worksheet. Your cell ranges should start with row 2 so as to exclude the row 1 variable names (an exception to the usual practice in this book). Remember to press the **Enter key** while holding down the **Control** and **Shift keys** (or the **Command key** on a Mac) to enter the array formula as discussed in Appendix Section B.3.

Read the **SHORT TAKES** for Chapter 14 for an explanation of the formulas found in the COMPUTE worksheet (shown in the **COMPUTE_FORMULAS worksheet**). If you use an Excel version that is older than Excel 2010, use the same-name worksheets in the **Multiple Regression 2007 workbook**.

Analysis ToolPak Use Regression.

For the example, open to the **DATA worksheet** of the **OmniPower workbook** and:

1. Select **Data → Data Analysis**.
2. In the Data Analysis dialog box, select **Regression** from the **Analysis Tools** list and then click **OK**.
3. Enter **A1:A35** as the **Input Y Range** and enter **B1:C35** as the **Input X Range**.
4. Check **Labels** and check **Confidence Level** and enter **95** in its box.
5. Click **New Worksheet Ply**.
6. Click **OK**.



Predicting the Dependent Variable Y

Key Technique Use the **MMULT** array function and the **T.INV.2T** function to help compute intermediate values that determine the confidence interval estimate and prediction interval.

Example Compute the Figure 14.3 confidence interval estimate and prediction interval for the OmniPower sales data shown on page 548.

PHStat Use the *PHStat* “Interpreting the Regression Coefficients” instructions but replace step 6 with the following steps 6 through 8:

6. Check **Confidence Interval Estimate & Prediction Interval** and enter **95** as the percentage for **Confidence level for intervals**.
 7. Enter a **Title** and click **OK**.
 8. In the new worksheet, enter **79** in cell **B6** and enter **400** in cell **B7**.
- These steps create a new worksheet that is discussed in the following *In-Depth Excel* instructions.

In-Depth Excel Use the **CIEandPI worksheet** of the **Multiple Regression workbook** as a template.

The worksheet already contains the data and formulas for the example. The worksheet uses the **MMULT** function (see Appendix Section F.4) in several array formulas that perform matrix operations.

Modifying this worksheet for other models with more than two independent variables requires knowledge that is beyond the scope of this book. For other models with two independent variables, first paste the data for those variables into columns B and C of the **MRArray worksheet** and adjust the number of entries in column A (all of which are **1**). Then, adjust the COMPUTE worksheet to reflect the new regression data, using the *In-Depth Excel* “Interpreting the Regression Coefficients” instructions. Finally, open to the **CIEandPI** worksheet and edit the array formula in cell range **B9:D11** and the labels in cells **A6** and **A7**.

Read the **SHORT TAKES** for Chapter 14 for an explanation of the formulas found in the **CIEandPI** worksheet (shown in the **CIEandPI_FORMULAS worksheet**). If you use an Excel version that is older than Excel 2010, use the **CIEandPI** worksheet in the **Multiple Regression 2007 workbook**.

EG14.2 r^2 , ADJUSTED r^2 , and the OVERALL F TEST

The coefficient of multiple determination, r^2 , the adjusted r^2 , and the overall *F* test are all computed as part of creating the multiple regression results worksheet using the Section EG14.1 instructions. If you use either the *PHStat* or *In-Depth Excel* instructions, formulas are used to compute these results in the **COMPUTE worksheet**. Formulas in cells B5, B7, B13, C12, C13, D12, and E12 copy values computed by the array formula in cell range L2:N6. In cell F12, the expression **F.DIST.RT(F test statistic, 1, error degrees of freedom)** computes the *p*-value for the overall *F* test.

EG14.3 RESIDUAL ANALYSIS for the MULTIPLE REGRESSION MODEL

Key Technique Use arithmetic formulas and some results from the multiple regression COMPUTE worksheet to compute residuals.

Example Perform the residual analysis for the OmniPower sales data discussed in Section 14.3, starting on page 553.

PHStat Use the Section EG14.1 “Interpreting the Regression Coefficients” *PHStat* instructions. Modify step 5 by checking **Residuals Table** and **Residual Plots** in addition to checking **Regression Statistics Table** and **ANOVA and Coefficients Table**.

In-Depth Excel Use the **RESIDUALS worksheet** of the **Multiple Regression workbook** as a template. Then construct residual plots for the residuals and the predicted value of *Y* and for the residuals and each of the independent variables.

For the example, the **RESIDUALS** worksheet uses the OmniPower sales data already in the **MRData worksheet** to compute the residuals. To compute residuals for other data, first use the Section EG14.1 “Interpreting the Regression Coefficients” *In-Depth Excel* instructions to modify the **MRData** and **COMPUTE** worksheets. Then, open to the **RESIDUALS** worksheet and:

1. If the number of independent variables is greater than 2, select column D, right-click, and click **Insert** from the shortcut menu. Repeat this step as many times as necessary to create the additional columns to hold all the *X* variables.
2. Paste the data for the *X* variables into columns, starting with column B.
3. Paste *Y* values in column E (or in the second-to-last column if there are more than two *X* variables).
4. For sample sizes smaller than 34, delete the extra rows. For sample sizes greater than 34, copy the predicted *Y* and residuals formulas down through the row containing the last pair of *X* and *Y* values. Also, add the new observation numbers in column A.

To construct the residual plots, open to the **RESIDUALS** worksheet and select pairs of columns and then use the Section EG2.5 *In-Depth Excel* “The Scatter Plot” instructions. (If you forgot to select the columns, Excel will construct a meaningless plot of all of the data in the **RESIDUALS** worksheet.) For example, to construct the residual plot for the residuals and the predicted value of *Y*, select columns D and F. (See Appendix Section B.7 for help in selecting a non-contiguous cell range.)

Read the SHORT TAKES for Chapter 14 for an explanation of the formulas found in the RESIDUALS worksheet (shown in the RESIDUALS_FORMULAS worksheet).

Analysis ToolPak Use the Section EG14.1 *Analysis ToolPak* instructions. Modify step 5 by checking **Residuals** and **Residual Plots** before clicking **New Worksheet Ply** and then **OK**. The **Residuals Plots** option constructs residual plots only for each independent variable. To construct a plot of the residuals and the predicted value of Y , select the predicted and residuals cells (in the RESIDUAL OUTPUT area of the regression results worksheet) and then apply the Section EG2.5 *In-Depth Excel* “The Scatter Plot” instructions.

EG14.4 INFERENCES CONCERNING the POPULATION REGRESSION COEFFICIENTS

The regression results worksheets created by using the Section EG14.1 instructions include the information needed to make the inferences discussed in Section 14.4.

EG14.5 TESTING PORTIONS of the MULTIPLE REGRESSION MODEL

Key Technique Adapt the Section EG14.1 “Interpreting the Regression Coefficients” instructions and the Section EG13.2 instructions to develop the regression analyses needed.

Example Test portions of the multiple regression model for the OmniPower sales data as discussed in Section 14.5, starting on page 558.

PHStat Use the Section EG14.1 *PHStat* “Interpreting the Regression Coefficients” instructions but modify step 6 by checking **Coefficients of Partial Determination** before you click **OK**.

In-Depth Excel Use one of the **CPD worksheets** of the **Multiple Regression workbook** as a template.

For the example, the **CPD_2 worksheet** already contains the data to compute the coefficients of partial determination. For other problems, you use a two-step process to compute the coefficients of partial determination. You first use the Section EG14.1 and the Section EG13.2 *In-Depth Excel* instructions to create all possible regression results worksheets in a copy of the **Multiple Regression workbook**. For example, if you have two independent variables, you perform three regression analyses: Y with X_1 and X_2 , Y with X_1 , and Y with X_2 , to create three regression results worksheets. Then, you open to the **CPD worksheet** for the number of independent variables (**CPD_2**, **CPD_3**, and **CPD_4 worksheets** are included) and follow the italicized instructions to copy and **Paste Special Values** (see Appendix Section B.4) from the regression results worksheets.

EG14.6 USING DUMMY VARIABLES and INTERACTION TERMS in REGRESSION MODELS

Dummy Variables

Key Technique Use **Find and Replace** to create a dummy variable from a two-level categorical variable. Before using **Find**

and Replace, copy and paste the categorical values to another column in order to preserve the original values.

Example From the two-level categorical variable Fireplace, create the dummy variable named FireplaceCoded that is used in the Figure 14.10 regression model on page 564.

In-Depth Excel For the example, open to the **DATA worksheet** of the **SilverSpring workbook** and:

1. Copy and paste the **Fireplace** values in column I to **column J** (the first empty column). Enter **FireplaceCoded** in **cell J1**.
2. Select **column J**.
3. Press **Ctrl+H** (the keyboard shortcut for **Find and Replace**). In the Find and Replace dialog box:
4. Enter **Yes** in the **Find what** box and enter **1** in the **Replace with** box.
5. Click **Replace All**. If a message box to confirm the replacement appears, click **OK** to continue.
6. Enter **No** in the **Find what** box and enter **0** in the **Replace with** box.
7. Click **Replace All**. If a message box to confirm the replacement appears, click **OK** to continue.
8. Click **Close**.

Categorical variables that have more than two levels require the use of formulas in multiple columns. For example, to create the dummy variables for Example 14.3 on page 565, two columns are needed. Assume that the three-level categorical variable mentioned in the example is in Column D of the opened worksheet. A first new column that contains formulas in the form $=\text{IF}(\text{column D cell} = \text{first level}, 1, 0)$ and a second new column that contains formulas in the form $=\text{IF}(\text{column D cell} = \text{second level}, 1, 0)$ would properly create the two dummy variables that the example requires.

Interactions

To create an interaction term, add a column of formulas that multiply one independent variable by another. For example, if the first independent variable appeared in column B and the second independent variable appeared in column C, enter the formula $=\text{B2} * \text{C2}$ in the row 2 cell of an empty new column and then copy the formula down through all rows of data to create the interaction.

EG14.7 LOGISTIC REGRESSION

Key Technique Use an automated process that incorporates the use of the Solver add-in to develop a logistic regression analysis model.

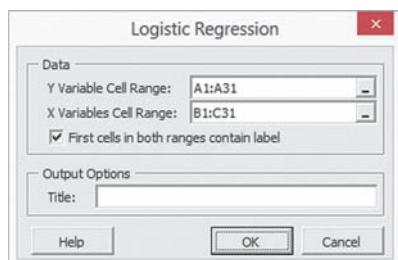
Example Develop the Figure 14.14 logistic regression model for the credit card pilot study data shown on page 575.

PHStat Use **Logistic Regression**.

For the example, open to the **DATA worksheet** of the **CardStudy workbook**. Select **PHStat → Regression → Logistic Regression**, and in the procedure’s dialog box (shown on page 592):

1. Enter **A1:A31** as the **Y Variable Cell Range**.
2. Enter **B1:C31** as the **X Variables Cell Range**.

3. Check **First cells in both ranges contain label**.
4. Enter a **Title** and click **OK**.



If the Solver add-in is not installed (see Appendix Section D.6), PHStat will display an error message instead of the Logistic Regression dialog box. The COMPUTE worksheet created contains a number of columns not shown in Figure 14.14 that contain supporting data.

In-Depth Excel Use the **Logistic Regression add-in workbook**. This workbook requires that the Solver add-in be installed (see Appendix Section D.6).

For the example, first open to the **DATA worksheet** of the **CardStudy workbook**. Then open the **Logistic Regression add-in workbook**. When this workbook opens properly, it adds a **Logistic Add-in** menu in either the Add-ins tab (Microsoft Windows) or the Apple menu bar (OS X). Select **Logistic Add-**

in → Logistic Regression from either the Add-ins tab or the Apple menu bar. In the Logistic Regression dialog box (shown below):

1. Enter **A1:A31** as the **Y Variable Cell Range**.
2. Enter **B1:C31** as the **X Variables Cell Range**.
3. Check **First cells in both ranges contain label**.
4. Enter a **Title** and click **OK**.



If the Solver add-in is not installed, you will see an error message in lieu of the Logistic Regression dialog box. This add-in workbook requires data workbooks to be in the **.xlsx** format and not the older **.xls** format.

EG14.8 INFLUENCE ANALYSIS

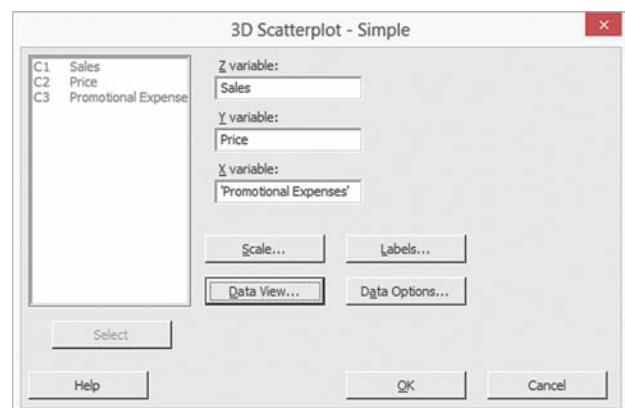
There are no Excel Guide instructions for this section.

CHAPTER 14 MINITAB GUIDE

MG14.1 DEVELOPING a MULTIPLE REGRESSION MODEL

Use **3D Scatterplot** to create a three-dimensional plot for the special case of a regression model that contains two independent variables. For example, to create the Figure 14.1 plot on page 545 for the OmniPower sales data, open the **OmniPower worksheet**. Select **Graph → 3D Scatterplot**. In the 3D Scatterplots dialog box, click **Simple** and then click **OK**. In the 3D Scatterplot - Simple dialog box (shown in right column):

1. Double-click **C1 Sales** in the variables list to add **Sales** to the **Z variable** box.
2. Double-click **C2 Price** in the variables list to add **Price** to the **Y variable** box.
3. Double-click **C3 Promotional Expenses** in the variables list to add '**Promotional Expenses**' to the **X variable** box.
4. Click **Data View**.



In the 3D Scatterplot - Data View dialog box:

5. Check **Symbols and Project lines**.
6. Click **OK**.
7. Back in the 3D Scatterplot - Simple dialog box, click **OK**.

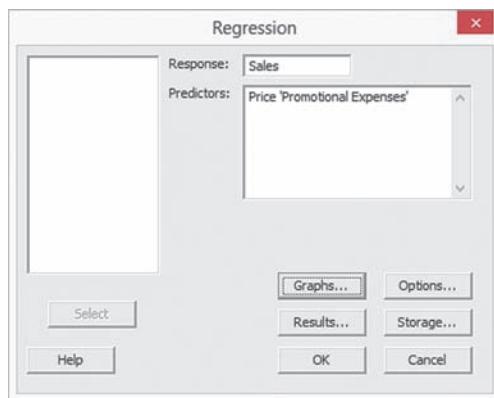
Rotate the scatter plot using the icons to rotate the *X*, *Y*, and *Z* axes in the 3D Graph Tools toolbar. Select **Tools → Toolbars → 3D Graph Tools** if this toolbar is not visible in the Minitab window.

The right scatter plot in Figure 14.1 was rotated clockwise about 90 degrees around the Z axis and was slightly rotated about the two other axes.

Interpreting the Regression coefficients

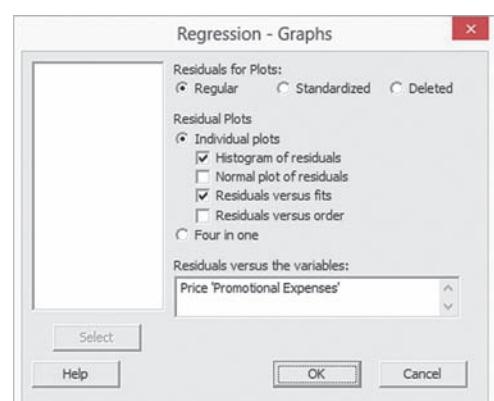
Use **Regression** to perform a multiple regression analysis. For example, to perform the Figure 14.2 analysis of the OmniPower sales data on page 546, open to the **OmniPower worksheet**. Select **Stat → Regression → Regression**. In the Regression dialog box (shown below):

1. Double-click **C1 Sales** in the variables list to add **Sales** to the **Response** box.
2. Double-click **C2 Price** in the variables list to add **Price** to the **Predictors** box.
3. Double-click **C3 Promotional Expenses** in the variables list to add '**Promotional Expenses**' to the **Predictors** box.
4. Click **Graphs**.



In the Regression - Graphs dialog box (shown below):

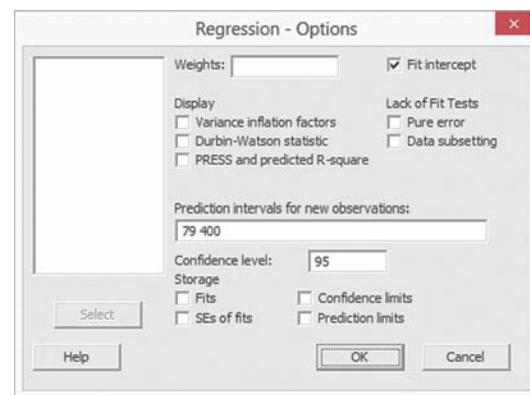
5. Click **Regular** and **Individual Plots**.
6. Check **Histogram of residuals** and **Residuals versus fits** and clear the other check boxes.
7. Click anywhere inside the **Residuals versus the variables** box.
8. Double-click **C2 Price** in the variables list to add **Price** in the **Residuals versus the variables** box.
9. Double-click **C3 Promotional Expenses** in the variables list to add '**Promotional Expenses**' in the **Residuals versus the variables** box.
10. Click **OK**.



11. Back in the Regression dialog box, click **Results**.

In the Regression - Results dialog box (not shown):

12. Click **In addition, the full table of fits and residuals** and then click **OK**.
 13. Back in the Regression dialog box, click **Options**.
- In the Regression - Options dialog box (shown below):
14. Check **Fit Intercept**.
 15. Clear all the **Display** and **Lack of Fit Test** check boxes.
 16. Enter **79** and **400** in the **Prediction intervals for new observations** box.
 17. Enter **95** in the **Confidence level** box.
 18. Click **OK**.



19. Back in the Regression dialog box, click **OK**.

The results in the Session Window will include additional items that are not shown in Figure 14.2.

Predicting the Dependent Variable Y

The regression results created by using the Section MG14.1 instructions include the confidence interval estimation and prediction interval. Figure 14.3 on page 548 shows these items for the OmniPower sales data.

MG14.2 r^2 , ADJUSTED r^2 , and the OVERALL F TEST

The coefficient of multiple determination, r^2 , the adjusted r^2 , and the overall F test are all computed as part of creating the multiple regression results using the Section MG14.1 instructions.

MG14.3 RESIDUAL ANALYSIS for the MULTIPLE REGRESSION MODEL

The regression results created by using the MG14.1 instructions include a residual analysis.

MG14.4 INFERENCES CONCERNING the POPULATION REGRESSION COEFFICIENTS

The regression results created by using the MG14.1 instructions include the information needed to make the inferences discussed in Section 14.4.

MG14.5 TESTING PORTIONS of the MULTIPLE REGRESSION MODEL

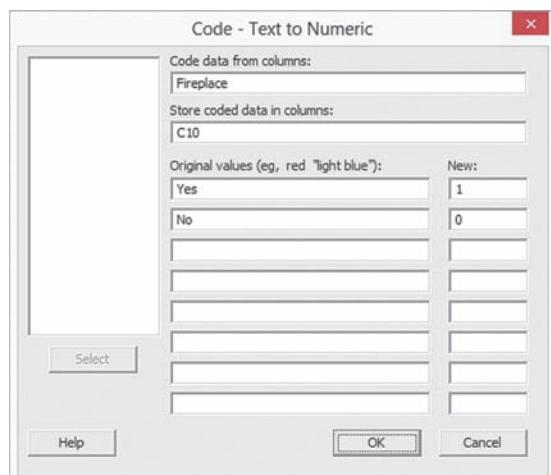
You compute the coefficients of partial determination by using a two-step process. You first use the Section MG14.1 instructions to create all possible regression results in the same project file. For example, if you have two independent variables, you perform three regression analyses— Y with X_1 and X_2 , Y with X_1 , and Y with X_2 —to create three sets of regression results. With those results you can then compute the partial F test and the coefficients of partial determination using the instructions in Section 14.5.

MG14.6 USING DUMMY VARIABLES and INTERACTION TERMS in REGRESSION MODELS

Dummy Variables

Use **Text to Numeric** to create a dummy variable. For example, to create from the categorical variable Fireplace the dummy variable named Fireplace Coded that is used in the Figure 14.10 regression model on page 564, open to the **SilverSpring worksheet**. Select **Data → Code → Text to Numeric**. In the Code - Text to Numeric dialog box (shown below):

1. Double-click **C9 Fireplace** in the variables list to add **Fireplace** to the **Code data from columns** box and press **Tab**.
2. Enter **C10** in the **Store coded data in columns** box and press **Tab**. (Column C10 is the first empty column in the worksheet.)
3. In the first row, enter **Yes** in the **Original Values** (eg, red "light blue") box and enter **1** in the **New** box.
4. In the second row, enter **No** in the **Original Values** (eg, red "light blue") box and enter **0** in the **New** box.
5. Click **OK**.

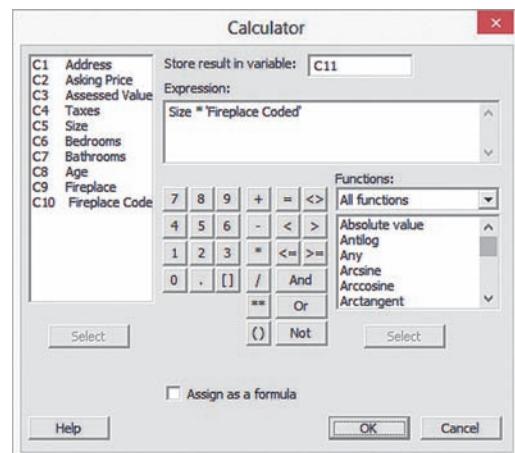


6. Enter **Fireplace Coded** as the name of column **C10**.

Interactions

Use **Calculator** to add a new column that contains the product of multiplying one independent variable by another to create an interaction term. For example, to create an interaction term of size and the dummy variable FireplaceCoded that is used in the Figure 14.11 regression model on page 566, open to the **SilverSpring worksheet**. Use the "Dummy Variables" instructions in the preceding part to create the **Fireplace Coded** column in the worksheet. Select **Calc → Calculator**. In the Calculator dialog box (shown below):

1. Enter **C11** in the **Store result in variable** box and press **Tab**.
2. Enter **Size * 'Fireplace Coded'** in the **Expression** box.
3. Click **OK**.



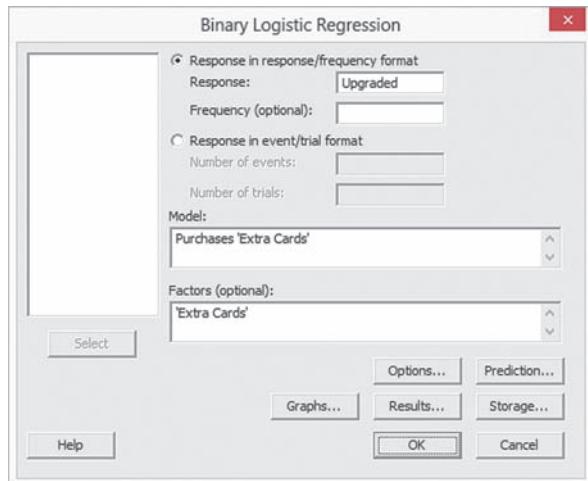
4. Enter **Size*Fireplace** as the name for column **C11**.

MG14.7 LOGISTIC REGRESSION

Use **Binary Logistic Regression** to perform a logistic regression. For example, to perform the Figure 14.14 logistic regression analysis shown on page 575, open to **CardStudy worksheet**. Select **Stat → Regression → Binary Logistic Regression**. In the Binary Logistic Regression dialog box (shown on page 595):

1. Click **Response in response/frequency format** and press **Tab**.
2. Double-click **C1 Upgraded** in the variables list to add **Upgraded** in the **Response** box.
3. Click inside the **Model** box.
4. Double-click **C2 Purchases** in the variables list to add **Purchases** to the **Model** box.
5. Double-click **C3 Extra Cards** in the variables list to add **'Extra Cards'** to the **Model** box and press **Tab**.
6. Double-click **C3 Extra Cards** in the variables list to add **'Extra Cards'** to the **Factors** box (because **Extra Cards** is a categorical variable).

7. Click OK.



MG14.8 INFLUENCE ANALYSIS

Use **Regression**. Use the Section MG14.1 “Interpreting the Regression Coefficients” instructions, replacing step 19 of those instructions with the steps 19 through 22 listed below.

For example, to perform the Figure 14.15 analysis of the OmniPower sales data shown on page 578, replace step 19 with these steps 19 through 22:

19. Back in the Regression dialog box, click **Storage**.

In the Regression - Storage dialog box:

20. Check **Deleted t residuals**, **Hi (leverages)**, and **Cook's distance**.

21. Click **OK**.

22. Back in the Regression dialog box, click **OK**.

CHAPTER 15

Multiple Regression Model Building

CONTENTS

- 15.1 The Quadratic Regression Model
- 15.2 Using Transformations in Regression Models
- 15.3 Collinearity
- 15.4 Model Building
- Steps for Successful Model Building**
- 15.5 Pitfalls in Multiple Regression and Ethical Issues

USING STATISTICS: Valuing Parsimony at WSTA-TV, Revisited

CHAPTER 15 EXCEL GUIDE

CHAPTER 15 MINITAB GUIDE

OBJECTIVES

- To use quadratic terms in a regression model
- To use transformed variables in a regression model
- To measure the correlation among independent variables
- To build a regression model using either the stepwise or best-subsets approach
- To avoid the pitfalls involved in developing a multiple regression model

USING STATISTICS

Valuing Parsimony at WSTA-TV

Your job as the broadcast operations manager at local station WSTA-TV has proven more challenging of late, as you adjust to initiatives that have been announced by the station's new owners. To meet the new business objective of reducing expenses by 8% during the next fiscal year, you seek to investigate ways to reduce unnecessary labor expenses associated with the staff of graphic artists employed by the station. Currently, these graphic artists receive hourly pay for a significant number of *standby hours*, hours for which they are present at the station but not assigned any specific task to do.

You believe that an appropriate model will help you to predict the number of future standby hours, identify the root causes of excessive numbers of standby hours, and allow you to reduce the total number of future standby hours. You plan to first collect weekly data for the number of standby hours and these four variables: the number of graphic artists present, the number of remote hours, the number of Dubner hours, and the total labor hours. Then, you seek to build a multiple regression model that will help determine which variables most heavily affect standby hours.

How do you build the model that has the most appropriate mix of independent variables? Are there statistical techniques that can help you identify a "best" model without having to consider all possible models?



Alamy

Chapter 14 discussed multiple regression models with two independent variables. This chapter considers regression models that contain more than two independent variables. The chapter discusses model-building concepts that will help to develop the best model when confronted with a set of data that has many independent variables, such as the data to be collected at WSTA-TV. These concepts include quadratic independent variables, transformations of the dependent or independent variables, stepwise regression, and best-subsets regression.

15.1 The Quadratic Regression Model

The simple regression model discussed in Chapter 13 and the multiple regression model discussed in Chapter 14 assume that the relationship between Y and each independent variable is linear. However, in Section 13.1, several different types of nonlinear relationships between variables were introduced. One of the most common nonlinear relationships is a quadratic, or curvilinear, relationship between two variables in which Y increases (or decreases) at a changing rate for various values of X (see Figure 13.1, Panels C through E, on page 492). You can use the quadratic regression model defined in Equation (15.1) to analyze this type of relationship between X and Y .

QUADRATIC REGRESSION MODEL

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i \quad (15.1)$$

where

β_0 = Y intercept

β_1 = coefficient of the linear effect on Y

β_2 = coefficient of the quadratic effect on Y

ε_i = random error in Y for observation i

This **quadratic regression model** is similar to the multiple regression model with two independent variables [see Equation (14.2) on page 545] except that the second independent variable, the **quadratic term**, is the square of the first independent variable. Once again, you use the least-squares method to compute sample regression coefficients (b_0 , b_1 , and b_2) as estimates of the population parameters (β_0 , β_1 , and β_2). Equation (15.2) defines the regression equation for the quadratic model with an independent variable (X_1) and a dependent variable (Y).

Student Tip

A quadratic regression model is a curvilinear model that has an X term and an X squared term. Other curvilinear models can have additional X terms that might involve X cubed, X raised to the fourth power, and so on.

QUADRATIC REGRESSION EQUATION

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{1i}^2 \quad (15.2)$$

In Equation (15.2), the first regression coefficient, b_0 , represents the Y intercept; the second regression coefficient, b_1 , represents the linear effect; and the third regression coefficient, b_2 , represents the quadratic effect.

Finding the Regression Coefficients and Predicting Y

To illustrate the quadratic regression model, consider a study that examined the business problem facing a concrete supplier of how adding fly ash affects the strength of concrete. (Fly ash is an inexpensive industrial waste by-product that can be used as a substitute for Portland cement, a more expensive ingredient of concrete.) Batches of concrete were prepared in which

the percentage of fly ash ranged from 0% to 60%. Data were collected from a sample of 18 batches and organized and stored in [FlyAsh](#). Table 15.1 summarizes the results.

TABLE 15.1

Fly Ash Percentage and Strength of 18 Batches of 28-Day-Old Concrete

Fly Ash %	Strength (psi)	Fly Ash %	Strength (psi)
0	4,779	40	5,995
0	4,706	40	5,628
0	4,350	40	5,897
20	5,189	50	5,746
20	5,140	50	5,719
20	4,976	50	5,782
30	5,110	60	4,895
30	5,685	60	5,030
30	5,618	60	4,648

By creating the scatter plot in Figure 15.1 to visualize these data, you will be better able to select the proper model for expressing the relationship between fly ash percentage and strength.

FIGURE 15.1

Scatter plot of fly ash percentage (X) and strength (Y)

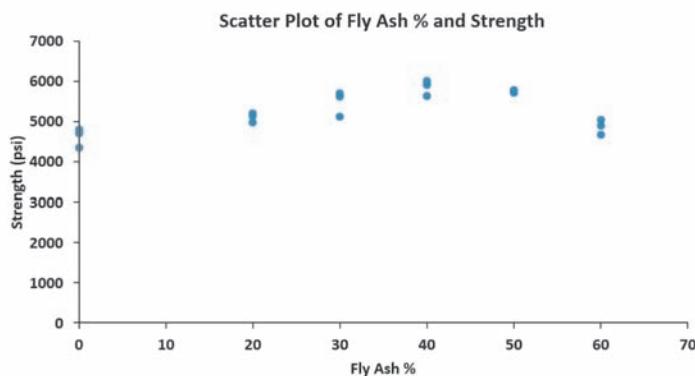


Figure 15.1 indicates an initial increase in the strength of the concrete as the percentage of fly ash increases. The strength appears to level off and then drop after achieving maximum strength at about 40% fly ash. Strength for 50% fly ash is slightly below strength at 40%, but strength at 60% fly ash is substantially below strength at 50%. Therefore, you should fit a quadratic model, not a linear model, to estimate strength based on fly ash percentage.

Figure 15.2 shows Excel and Minitab results for these data.
From Figure 15.2,

$$b_0 = 4,486.3611 \quad b_1 = 63.0052 \quad b_2 = -0.8765$$

FIGURE 15.2

Excel and Minitab regression results for the concrete strength data

Concrete Strength Analysis						
Regression Statistics						
Multiple R	0.8053					
R Square	0.6485					
Adjusted R Square	0.6016					
Standard Error	312.1129					
Observations	18					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	2695473.4897	1347736.745	13.8351	0.0004	
Residual	15	1461217.0103	97414.4674			
Total	17	4156690.5000				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	4486.3611	174.7531	25.6726	0.0000	4113.8834	4858.8389
Fly Ash%	63.0052	12.3725	5.0923	0.0001	36.6338	89.3767
Fly Ash% ^2	-0.8765	0.1966	-4.4578	0.0005	-1.2955	-0.4574

Regression Analysis: Strength versus Fly Ash%, Fly Ash%^2						
The regression equation is Strength = 4486 + 63.0 Fly Ash% - 0.876 Fly Ash%^2						
Predictor	Coeff	SE Coef	T	P		
Constant	4486.4	174.8	25.67	0.000		
Fly Ash%	63.01	12.37	5.09	0.000		
Fly Ash%^2	-0.8765	0.1966	-4.46	0.000		
S = 312.113	R-Sq = 64.8%	R-Sq(adj) = 60.2%				
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	2	2695473	1347737	13.84	0.000	
Residual Error	15	1461217	97414			
Total	17	4156690				

Therefore, the quadratic regression equation is

$$\hat{Y}_i = 4,486.3611 + 63.0052X_{1i} - 0.8765X_{1i}^2$$

where

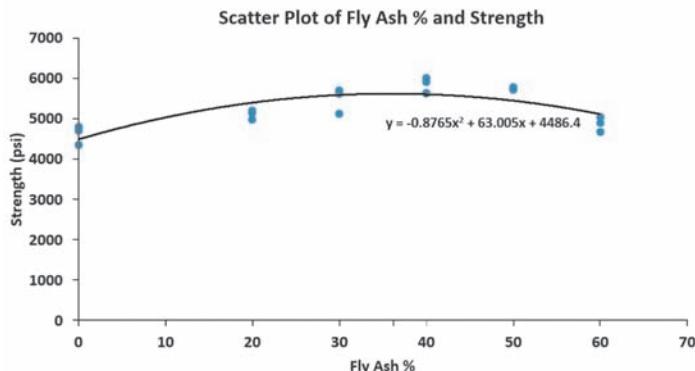
\hat{Y}_i = predicted strength for sample i

X_{1i} = percentage of fly ash for sample i

Figure 15.3 is a scatter plot of this quadratic regression equation that shows the fit of the quadratic regression model to the original data.

FIGURE 15.3

Scatter plot showing the quadratic relationship between fly ash percentage and strength for the concrete data



From the quadratic regression equation and Figure 15.3, the Y intercept ($b_0 = 4,486.3611$) is the predicted strength when the percentage of fly ash is 0. To interpret the coefficients b_1 and b_2 , observe that after an initial increase, strength decreases as fly ash percentage increases. This nonlinear relationship is further demonstrated by predicting the strength for fly ash percentages of 20, 40, and 60. Using the quadratic regression equation,

$$\hat{Y}_i = 4,486.3611 + 63.0052X_{1i} - 0.8765X_{1i}^2$$

for $X_{1i} = 20$,

$$\hat{Y}_i = 4,486.3611 + 63.0052(20) - 0.8765(20)^2 = 5,395.865$$

for $X_{1i} = 40$,

$$\hat{Y}_i = 4,486.3611 + 63.0052(40) - 0.8765(40)^2 = 5,604.169$$

and for $X_{1i} = 60$,

$$\hat{Y}_i = 4,486.3611 + 63.0052(60) - 0.8765(60)^2 = 5,111.273$$

Thus, the predicted concrete strength for 40% fly ash is 208.304 psi above the predicted strength for 20% fly ash, but the predicted strength for 60% fly ash is 492.896 psi below the predicted strength for 40% fly ash. The concrete supplier should consider using a fly ash percentage of 40% and not using fly ash percentages of 20% or 60% because those percentages lead to reduced concrete strength.

Student Tip

Remember that you are testing whether at least one independent variable has a linear relationship with the dependent variable. If you reject H_0 , you are *not* concluding that all the independent variables have a linear relationship with the dependent variable, only that *at least one* independent variable does.

Testing for the Significance of the Quadratic Model

After you calculate the quadratic regression equation, you can test whether there is a significant overall relationship between strength, Y , and fly ash percentage, X_1 . The null and alternative hypotheses are as follows:

$H_0: \beta_1 = \beta_2 = 0$ (There is no overall relationship between X_1 and Y .)

$H_1: \beta_1$ and/or $\beta_2 \neq 0$ (There is an overall relationship between X_1 and Y .)

Equation (14.6) on page 551 defines the overall F_{STAT} test statistic used for this test:

$$F_{STAT} = \frac{MSR}{MSE}$$

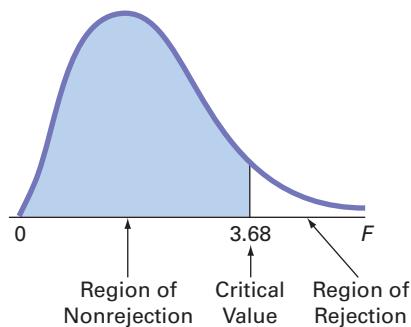
From the Figure 15.2 results on page 598,

$$F_{STAT} = \frac{MSR}{MSE} = \frac{1,347,736.745}{97,414.4674} = 13.8351$$

If you choose a level of significance of 0.05, from Table E.5, the critical value of the F distribution, with 2 and $18 - 2 - 1 = 15$ degrees of freedom, is 3.68 (see Figure 15.4). Because $F_{STAT} = 13.8351 > 3.68$, or because the p -value = 0.0004 < 0.05, you reject the null hypothesis (H_0) and conclude that there is a significant overall relationship between strength and fly ash percentage.

FIGURE 15.4

Testing for the existence of the overall relationship at the 0.05 level of significance, with 2 and 15 degrees of freedom



Testing the Quadratic Effect

In using a regression model to examine a relationship between two variables, you want to find not only the most accurate model but also the simplest model that expresses that relationship. Therefore, you need to examine whether there is a significant difference between the quadratic model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

and the linear model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

In Section 14.4, you used the t test to determine whether each independent variable makes a significant contribution to the regression model. To test the significance of the contribution of the quadratic effect, you use the following null and alternative hypotheses:

- H_0 : Including the quadratic effect does not significantly improve the model ($\beta_2 = 0$).
- H_1 : Including the quadratic effect significantly improves the model ($\beta_2 \neq 0$).

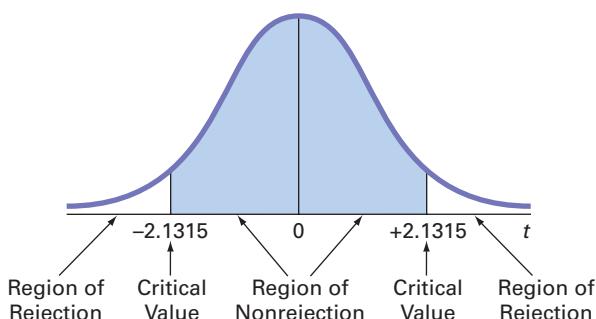
The standard error of each regression coefficient and its corresponding t_{STAT} test statistic are part of the regression results (see Figure 15.2 on page 598). Equation (14.7) on page 555 defines the t_{STAT} test statistic:

$$\begin{aligned} t_{STAT} &= \frac{b_2 - \beta_2}{S_{b_2}} \\ &= \frac{-0.8765 - 0}{0.1966} = -4.4578 \end{aligned}$$

If you select the 0.05 level of significance, then from Table E.3, the critical values for the t distribution with 15 degrees of freedom are -2.1315 and $+2.1315$ (see Figure 15.5).

FIGURE 15.5

Testing for the contribution of the quadratic effect to a regression model at the 0.05 level of significance, with 15 degrees of freedom



Because $t_{STAT} = -4.4578 < -2.1315$ or because the p -value = 0.0005 < 0.05, you reject H_0 and conclude that the quadratic model is significantly better than the linear model for representing the relationship between strength and fly ash percentage.

Example 15.1 provides an additional illustration of a possible quadratic effect.

EXAMPLE 15.1

Studying the Quadratic Effect in a Multiple Regression Model

A real estate developer studying the business problem of estimating the consumption of heating oil by single-family houses has decided to examine the effect of atmospheric temperature and the amount of attic insulation on heating oil consumption. Data are collected from a random sample of 15 single-family houses. The data are organized and stored in **HeatingOil**. Figure 15.6 shows the regression results for a multiple regression model using the two independent variables: atmospheric temperature and attic insulation.

FIGURE 15.6

Excel and Minitab results for the multiple linear regression model predicting monthly consumption of heating oil

A	B	C	D	E	F	G
1 Heating Oil Consumption Analysis						
2						
3 Regression Statistics						
4 Multiple R	0.9827					
5 R Square	0.9656					
6 Adjusted R Square	0.9599					
7 Standard Error	26.0138					
8 Observations	15					
9						
10 ANOVA						
11 df	SS	MS	F	Significance F		
12 Regression	2	228014.6263	114007.3132	168.4712	0.0000	
13 Residual	12	8120.6030	676.7169			
14 Total	14	236135.2293				
15						
16 Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	
17 Intercept	562.1510	21.0931	26.6509	0.0000	516.1931	608.1089
18 Temperature	-5.4366	0.3362	-16.1699	0.0000	-6.1691	-4.7040
19 Insulation	-20.0123	2.3425	-8.5431	0.0000	-25.1162	-14.9084

Regression Analysis: Gallons versus Temperature, Insulation

The regression equation is
Gallons = 562 - 5.44 Temperature - 20.0 Insulation

Predictor	Coef	SE Coef	T	P
Constant	562.15	21.09	26.65	0.000
Temperature	-5.4366	0.3362	-16.17	0.000
Insulation	-20.012	2.343	-8.54	0.000

S = 26.0138 R-Sq = 96.6% R-Sq(adj) = 96.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	228015	114007	168.47	0.000
Residual Error	12	8121	677		
Total	14	236135			

The residual plot for attic insulation (not shown here) contained some evidence of a quadratic effect. Thus, the real estate developer reanalyzed the data by adding a quadratic term for attic insulation to the multiple regression model. At the 0.05 level of significance, is there evidence of a significant quadratic effect for attic insulation?

(continued)

SOLUTION Figure 15.7 shows the results for this regression model.

FIGURE 15.7

Excel and Minitab results for the multiple regression model with a quadratic term for attic insulation

A	B	C	D	E	F	G
1 Quadratic Effect for Insulation Variable?						
2						
3 Regression Statistics						
4 Multiple R	0.9862					
5 R Square	0.9725					
6 Adjusted R Square	0.9650					
7 Standard Error	24.2938					
8 Observations	15					
9						
10 ANOVA						
11	df	SS	MS	F	Significance F	
12 Regression	3	229643.1645	76547.7215	129.7006	0.0000	
13 Residual	11	6492.0649	590.1877			
14 Total	14	236135.2293				
15						
16	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17 Intercept	624.5864	42.4352	14.7186	0.0000	531.1872	717.9856
18 Temperature	-5.3626	0.3171	-16.9099	0.0000	-6.0606	-4.6646
19 Insulation	-44.5868	14.9547	-2.9815	0.0125	-77.5019	-11.6717
20 Insulation^2	1.8667	1.1238	1.6611	0.1249	-0.6067	4.3401

Regression Analysis: Gallons versus Temperature, Insulation, ...						
The regression equation is Gallons = 625 - 5.36 Temperature - 44.6 Insulation + 1.87 Insulation^2						
Predictor	Coef	SE Coef	T	P		
Constant	624.59	42.44	14.72	0.000		
Temperature	-5.3626	0.3171	-16.91	0.000		
Insulation	-44.59	14.95	-2.98	0.012		
Insulation^2	1.867	1.124	1.66	0.125		
S = 24.2938 R-Sq = 97.3% R-Sq(adj) = 96.5%						
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	3	229643	76548	129.70	0.000	
Residual Error	11	6492	590			
Total	14	236135				

The multiple regression equation is

$$\hat{Y}_i = 624.5864 - 5.3626X_{1i} - 44.5868X_{2i} + 1.8667X_{2i}^2$$

To test for the significance of the quadratic effect:

H_0 : Including the quadratic effect of insulation does not significantly improve the model ($\beta_3 = 0$).

H_1 : Including the quadratic effect of insulation significantly improves the model ($\beta_3 \neq 0$).

From Figure 15.7 and Table E.3 with $15 - 3 - 1 = 11$ degrees of freedom, $-2.2010 < t_{STAT} = 1.6611 < 2.2010$ (or the p -value = 0.1249 > 0.05). Therefore, you do not reject the null hypothesis. You conclude that there is insufficient evidence that the quadratic effect for attic insulation is different from zero. In the interest of keeping the model as simple as possible, you should use the multiple regression equation shown in Figure 15.6:

$$\hat{Y}_i = 562.1510 - 5.4366X_{1i} - 20.0123X_{2i}$$

Student Tip

Remember that r^2 in multiple regression represents the proportion of the variation in the dependent variable Y that is explained by *all* the independent X variables included in the model. So, in this case of quadratic regression, r^2 represents the proportion of the variation in the dependent variable Y that is explained by the linear term and the quadratic term.

The Coefficient of Multiple Determination

In the multiple regression model, the coefficient of multiple determination, r^2 (see Section 14.2), represents the proportion of variation in Y that is explained by variation in the independent variables. Consider the quadratic regression model you used to predict the strength of concrete using fly ash and fly ash squared. You compute r^2 by using Equation (14.4) on page 550:

$$r^2 = \frac{SSR}{SST}$$

From Figure 15.2 on page 598,

$$SSR = 2,695,473.4897 \quad SST = 4,156,690.5$$

Thus,

$$r^2 = \frac{SSR}{SST} = \frac{2,695,473.4897}{4,156,690.5} = 0.6485$$

This coefficient of multiple determination indicates that 64.85% of the variation in strength is explained by the quadratic relationship between strength and the percentage of fly ash. You should also compute r_{adj}^2 to account for the number of independent variables and the sample size. In the quadratic regression model, $k = 2$ because there are two independent variables, X_1 and X_1^2 . Thus, using Equation (14.5) on page 550,

$$\begin{aligned} r_{adj}^2 &= 1 - \left[(1 - r^2) \frac{(n - 1)}{(n - k - 1)} \right] \\ &= 1 - \left[(1 - 0.6485) \frac{17}{15} \right] \\ &= 1 - 0.3984 \\ &= 0.6016 \end{aligned}$$

Problems for Section 15.1

LEARNING THE BASICS

- 15.1** The following is the quadratic regression equation for a sample of $n = 25$:

$$\hat{Y}_i = 5 + 3X_{1i} + 1.5X_{1i}^2$$

- Predict Y for $X_1 = 2$.
- Suppose that the computed t_{STAT} test statistic for the quadratic regression coefficient is 2.35. At the 0.05 level of significance, is there evidence that the quadratic model is better than the linear model?
- Suppose that the computed t_{STAT} test statistic for the quadratic regression coefficient is 1.17. At the 0.05 level of significance, is there evidence that the quadratic model is better than the linear model?
- Suppose the regression coefficient for the linear effect is -3.0 . Predict Y for $X_1 = 2$.

APPLYING THE CONCEPTS

- 15.2** Businesses actively recruit business students with well-developed higher-order cognitive skills (HOCS) such as problem identification, analytical reasoning, and content integration skills. Researchers conducted a study to see if improvement in students' HOCS was related to the students' GPA. (Data extracted from R. V. Bradley, C. S. Sankar, H. R. Clayton, V. W. Mbarika, and P. K. Raju, "A Study on the Impact of GPA on Perceived Improvement of Higher-Order Cognitive Skills," *Decision Sciences Journal of Innovative Education*, January 2007, 5(1), pp. 151–168.) The researchers conducted a study in which business students were taught using the case study method. Using data collected from 300 business students, the following quadratic regression equation was derived:

$$\text{HOCS} = -3.48 + 4.53(\text{GPA}) - 0.68(\text{GPA})^2$$

where the dependent variable HOCS measured the improvement in higher-order cognitive skills, with 1 being the lowest improvement in HOCS and 5 being the highest improvement in HOCS.

- Construct a table of predicted HOCS, using GPA equal to 2.0, 2.1, 2.2, ..., 4.0.
- Plot the values in the table constructed in (a), with GPA on the horizontal axis and predicted HOCS on the vertical axis.

- Discuss the curvilinear relationship between students' GPA and their predicted improvement in HOCS.
- The researchers reported that the model had an r^2 of 0.07 and an adjusted r^2 of 0.06. What does this tell you about the scatter of individual HOCS scores around the curvilinear relationship plotted in (b) and discussed in (c)?

- 15.3** A national chain of consumer electronics stores had the business objective of determining the effectiveness of newspaper advertising. To promote sales, the chain relies heavily on local newspaper advertising to support its modest exposure in nationwide television commercials. A sample of 20 cities with similar populations and monthly sales totals were assigned different newspaper advertising budgets for one month. The following table, stored in **Advertising**, summarizes the sales (in \$millions) and the newspaper advertising budgets (in \$thousands) observed during the study:

Sales (\$millions)	Newspaper Advertising (\$thousands)	Sales (\$millions)	Newspaper Advertising (\$thousands)
6.14	5	6.84	15
6.04	5	6.66	15
6.21	5	6.95	20
6.32	5	6.65	20
6.42	10	6.83	20
6.56	10	6.81	20
6.67	10	7.03	25
6.35	10	6.88	25
6.76	15	6.84	25
6.79	15	6.99	25

- Construct a scatter plot for newspaper advertising and sales.
- Fit a quadratic regression model and state the quadratic regression equation.
- Predict the mean monthly sales for a city with newspaper advertising of \$20,000.
- Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- At the 0.05 level of significance, is there a significant quadratic relationship between monthly sales and newspaper advertising?

- f. At the 0.05 level of significance, determine whether the quadratic model is a better fit than the linear model.
- g. Interpret the meaning of the coefficient of multiple determination.
- h. Compute the adjusted r^2 .
- i. What conclusions can you reach concerning the relationship between newspaper advertising and sales?

15.4 Is the number of calories in a beer related to the number of carbohydrates and/or the percentage of alcohol in the beer? Data concerning 152 of the best-selling domestic beers in the United States are stored in **DomesticBeer**. The values for three variables are included: the number of calories per 12 ounces, the alcohol percentage, and the number of carbohydrates (in grams) per 12 ounces. (Data extracted from www.beer100.com/beercalories.htm, March 21, 2013.)

- a. Perform a multiple linear regression analysis, using calories as the dependent variable and percentage alcohol and number of carbohydrates as the independent variables.
- b. Add quadratic terms for alcohol percentage and the number of carbohydrates.
- c. Which model is better, the one in (a) or (b)?
- d. What conclusions can you reach concerning the relationship between the number of calories in a beer and the alcohol percentage and number of carbohydrates?

15.5 In the production of printed circuit boards, errors in the alignment of electrical connections are a source of scrap. The data in the file **RegistrationError-HighCost** contains the registration error and the temperature used in the production of circuit boards in an experiment in which higher cost material was used. (Data extracted from C. Nachtsheim and B. Jones, "A Powerful Analytical Tool," *Six Sigma Forum Magazine*, August 2003, pp. 30–33.)

- a. Construct a scatter plot for temperature and registration error.
- b. Fit a quadratic regression model to predict registration error and state the quadratic regression equation.
- c. Perform a residual analysis on the results and determine whether the regression model is valid.
- d. At the 0.05 level of significance, is there a significant quadratic relationship between temperature and registration error?
- e. At the 0.05 level of significance, determine whether the quadratic model is a better fit than the linear model.
- f. Interpret the meaning of the coefficient of multiple determination.
- g. Compute the adjusted r^2 .
- h. What conclusions can you reach concerning the relationship between registration error and temperature?

 **15.6** A production manager wishes to examine the relationship between unit production (number of units produced) and associated costs (total cost). The file **CostEstimation** contains data for 10 months of production.

- a. Construct a scatter plot for unit production and total cost.
- b. Fit a quadratic regression model to predict total cost and state the quadratic regression equation.
- c. Predict the mean total cost when production is 145 units.
- d. Perform a residual analysis on the results and determine whether the regression model is valid.
- e. At the 0.05 level of significance, is there a significant quadratic relationship between monthly unit production and total cost?
- f. What is the p -value in (e)? Interpret its meaning.
- g. At the 0.05 level of significance, determine whether the quadratic model is a better fit than the linear model.
- h. What is the p -value in (g)? Interpret its meaning.
- i. Interpret the meaning of the coefficient of multiple determination.
- j. Compute the adjusted r^2 .
- k. What conclusions can you reach concerning the relationship between cost and unit production?

15.7 Researchers wanted to investigate the relationship between employment and accommodation capacity in the European travel and tourism industry. The file **EuroTourism** contains a sample of 27 European countries. Variables included are the number of jobs generated in the travel and tourism industry in 2012 and the number of establishments that provide overnight accommodation for tourists. (Data extracted from www.marketline.com.)

- a. Construct a scatter plot of the number of jobs generated in the travel and tourism industry in 2012 (Y) and the number of establishments that provide overnight accommodation for tourists (X).
- b. Fit a quadratic regression model to predict the number of jobs generated and state the quadratic regression equation.
- c. Predict the mean number of jobs generated in the travel and tourism industry for a country with 3,000 establishments that provide overnight accommodation for tourists.
- d. Perform a residual analysis on the results and determine whether the regression model is valid.
- e. At the 0.05 level of significance, is there a significant quadratic relationship between the number of jobs generated in the travel and tourism industry in 2012 and the number of establishments that provide overnight accommodation for tourists?
- f. What is the p -value in (e)? Interpret its meaning.
- g. At the 0.05 level of significance, determine whether the quadratic model is a better fit than the linear model.
- h. What is the p -value in (g)? Interpret its meaning.
- i. Interpret the meaning of the coefficient of multiple determination.
- j. Compute the adjusted r^2 .
- k. What conclusions can you reach concerning the relationship between the number of jobs generated in the travel and tourism industry in 2012 and the number of establishments that provide overnight accommodation for tourists?

15.2 Using Transformations in Regression Models

This section introduces regression models in which the independent variable, the dependent variable, or both are transformed in order to either overcome violations of the assumptions of regression or to make a model whose form is not linear into a linear model. Among the many transformations available (see reference 1) are the square-root transformation and transformations involving the common logarithm (base 10) and the natural logarithm (base e).¹

¹For more information on logarithms, see Appendix Section A.3.

Student Tip

The log (\log) of a number is the power to which 10 needs to be raised to equal that number. The natural log (\ln) of a number is the power to which e , Euler's number, needs to be raised to equal that number.

The Square-Root Transformation

The **square-root transformation** is often used to overcome violations of the equal-variance assumption as well as to transform a model whose form is not linear into a linear model. Equation (15.3) shows a regression model that uses a square-root transformation of the independent variable.

REGRESSION MODEL WITH A SQUARE-ROOT TRANSFORMATION

$$Y_i = \beta_0 + \beta_1 \sqrt{X_{1i}} + \varepsilon_i \quad (15.3)$$

Example 15.2 illustrates the use of a square-root transformation.

EXAMPLE 15.2

Given the following values for X and Y , use a square-root transformation for the X variable:

Using the Square-Root Transformation

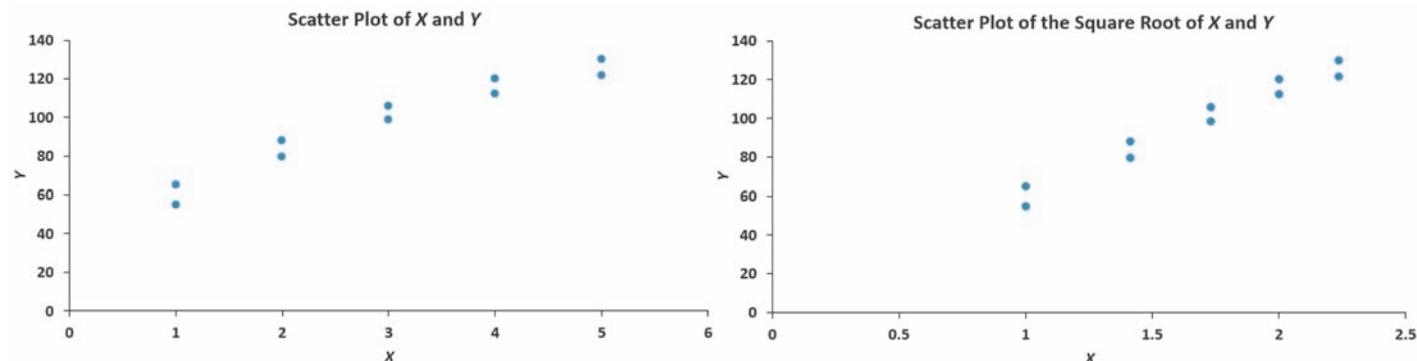
X	Y	X	Y
1	55.0	3	105.7
1	65.0	4	120.1
2	87.9	4	112.3
2	79.8	5	121.5
3	98.5	5	129.9

Construct a scatter plot for X and Y and for the square root of X and Y .

SOLUTION Figure 15.8 displays both scatter plots.

FIGURE 15.8

Example 15.2 scatter plots of X and Y and the square root of X and Y



You can see that the square-root transformation has transformed a nonlinear relationship into a linear relationship.

The Log Transformation

The **logarithmic transformation** is often used to overcome violations of the equal-variance assumption. You can also use the logarithmic transformation to change a nonlinear model into a linear model. Equation (15.4) shows a multiplicative model.

ORIGINAL MULTIPLICATIVE MODEL

$$Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} \varepsilon_i \quad (15.4)$$

By taking base 10 logarithms of both the dependent and independent variables, you can transform Equation (15.4) to the model shown in Equation (15.5).

TRANSFORMED MULTIPLICATIVE MODEL

$$\begin{aligned} \log Y_i &= \log(\beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} \varepsilon_i) \\ &= \log \beta_0 + \log(X_{1i}^{\beta_1}) + \log(X_{2i}^{\beta_2}) + \log \varepsilon_i \\ &= \log \beta_0 + \beta_1 \log X_{1i} + \beta_2 \log X_{2i} + \log \varepsilon_i \end{aligned} \quad (15.5)$$

Thus, Equation (15.5) is linear in the logarithms. Similarly, you can transform the exponential model shown in Equation (15.6) to a linear form by taking the natural logarithm of both sides of the equation. Equation (15.7) is the transformed model.

ORIGINAL EXPONENTIAL MODEL

$$Y_i = e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}} \varepsilon_i \quad (15.6)$$

TRANSFORMED EXPONENTIAL MODEL

$$\begin{aligned} \ln Y_i &= \ln(e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}} \varepsilon_i) \\ &= \ln(e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}}) + \ln \varepsilon_i \\ &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ln \varepsilon_i \end{aligned} \quad (15.7)$$

Example 15.3 illustrates the use of a natural log transformation.

EXAMPLE 15.3

Given the following values for X and Y , use a natural logarithm transformation for the Y variable:

Using the Natural Log Transformation

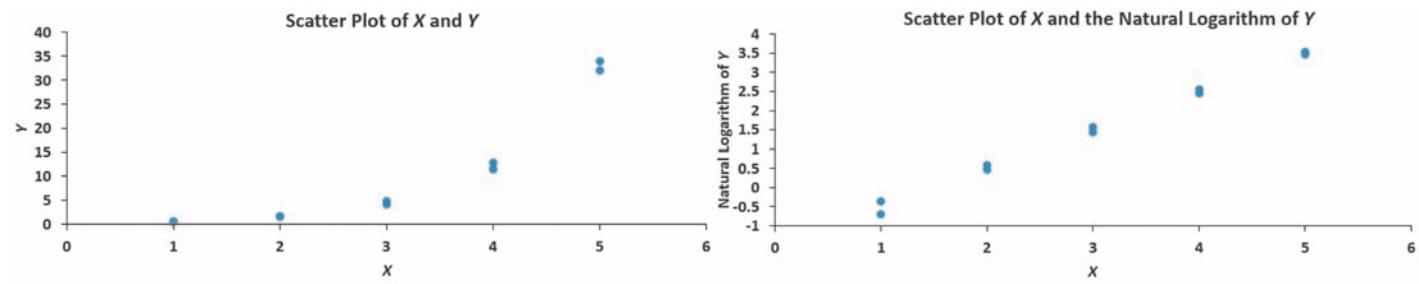
X	Y	X	Y
1	0.7	3	4.8
1	0.5	4	12.9
2	1.6	4	11.5
2	1.8	5	32.1
3	4.2	5	33.9

Construct a scatter plot for X and Y and for X and the natural logarithm of Y .

SOLUTION Figure 15.9 displays both scatter plots. The plots show that the natural logarithm transformation has changed a nonlinear relationship into a linear relationship.

FIGURE 15.9

Example 15.3 scatter plots of X and Y and X and the natural logarithm of Y



Problems for Section 15.2

LEARNING THE BASICS

15.8 Consider the following regression equation:

$$\log \hat{Y}_i = \log 3.07 + 0.9 \log X_{1i} + 1.41 \log X_{2i}$$

- Predict the value of Y when $X_1 = 8.5$ and $X_2 = 5.2$.
- Interpret the meaning of the regression coefficients b_0 , b_1 , and b_2 .

15.9 Consider the following regression equation:

$$\ln \hat{Y}_i = 4.62 + 0.5X_{1i} + 0.7X_{2i}$$

- Predict the value of Y when $X_1 = 8.5$ and $X_2 = 5.2$.
- Interpret the meaning of the regression coefficients b_0 , b_1 , and b_2 .

APPLYING THE CONCEPTS

SELF TEST **15.10** Using the data of Problem 15.4 on page 604, stored in **DomesticBeer**, perform a square-root transformation on each of the independent variables (percentage alcohol and number of carbohydrates). Using calories as the dependent variable and the transformed independent variables, perform a multiple regression analysis.

- State the regression equation.
- Perform a residual analysis of the results and determine whether the regression model is valid.
- At the 0.05 level of significance, is there a significant relationship between calories and the square root of the percentage of alcohol and the square root of the number of carbohydrates?
- Interpret the meaning of the coefficient of determination, r^2 , in this problem.
- Compute the adjusted r^2 .
- Compare your results with those in Problem 15.4. Which model is better? Why?

15.11 Using the data of Problem 15.4 on page 604, stored in **DomesticBeer**, perform a natural logarithmic transformation of the dependent variable (calories). Using the transformed dependent variable and the percentage of alcohol and the number of carbohydrates as the independent variables, perform a multiple regression analysis.

- State the regression equation.
- Perform a residual analysis of the results and determine whether the regression assumptions are valid.

- At the 0.05 level of significance, is there a significant relationship between the natural logarithm of calories and the percentage of alcohol and the number of carbohydrates?
- Interpret the meaning of the coefficient of determination, r^2 , in this problem.
- Compute the adjusted r^2 .
- Compare your results with those in Problems 15.4 and 15.10. Which model is best? Why?

15.12 Using the data of Problem 15.6 on page 604, stored in **CostEstimation**, perform a natural logarithm transformation of the dependent variable (total cost). Using the transformed dependent variable and the unit production as the independent variable, perform a regression analysis.

- State the regression equation.
- Predict the mean total cost when production is 145 units.
- Perform a residual analysis of the results and determine whether the regression assumptions are valid.
- At the 0.05 level of significance, is there a significant relationship between the natural logarithm of total cost and unit production?
- Interpret the meaning of the coefficient of determination, r^2 , in this problem.
- Compute the adjusted r^2 .
- Compare your results with those in Problem 15.6. Which model is better? Why?

15.13 Using the data of Problem 15.6 on page 604, stored in **CostEstimation**, perform a square-root transformation of the independent variable (unit production). Using total cost as the dependent variable and the transformed independent variable, perform a regression analysis.

- State the regression equation.
- Predict the mean total cost when production is 145 units.
- Perform a residual analysis of the results and determine whether the regression model is valid.
- At the 0.05 level of significance, is there a significant relationship between total cost and the square root of unit production?
- Interpret the meaning of the coefficient of determination, r^2 , in this problem.
- Compute the adjusted r^2 .
- Compare your results with those of Problems 15.6 and 15.12. Which model is best? Why?

15.3 Collinearity

One important problem in developing multiple regression models involves the possible **collinearity** of the independent variables. This condition refers to situations in which two or more of the independent variables are highly correlated with each other. In such situations, collinear variables do not provide unique information, and it becomes difficult to separate the effects of such variables on the dependent variable. When collinearity exists, the values of the regression coefficients for the correlated variables may fluctuate drastically, depending on which independent variables are included in the model.

One method of measuring collinearity is to determine the **variance inflationary factor (VIF)** for each independent variable. Equation (15.8) defines VIF_j , the variance inflationary factor for variable j .

VARIANCE INFLATIONARY FACTOR

$$VIF_j = \frac{1}{1 - R_j^2} \quad (15.8)$$

where R_j^2 is the coefficient of multiple determination for a regression model, using variable X_j as the dependent variable and all other X variables as independent variables.

If there are only two independent variables, R_1^2 is the coefficient of determination between X_1 and X_2 . It is identical to R_2^2 , which is the coefficient of determination between X_2 and X_1 . If there are three independent variables, then R_1^2 is the coefficient of multiple determination of X_1 with X_2 and X_3 ; R_2^2 is the coefficient of multiple determination of X_2 with X_1 and X_3 ; and R_3^2 is the coefficient of multiple determination of X_3 with X_1 and X_2 .

If a set of independent variables is uncorrelated, each VIF_j is equal to 1. If the set is highly correlated, then a VIF_j might even exceed 10. Marquardt (see reference 2) suggests that if VIF_j is greater than 10, there is too much correlation between the variable X_j and the other independent variables. However, other statisticians suggest a more conservative criterion. Snee (see reference 5) recommends using alternatives to least-squares regression if the maximum VIF_j exceeds 5.

You need to proceed with extreme caution when using a multiple regression model that has one or more large VIF values. You can use the model to predict values of the dependent variable *only* in the case where the values of the independent variables used in the prediction are in the relevant range of the values in the data set. However, you cannot extrapolate to values of the independent variables not observed in the sample data. And because the independent variables contain overlapping information, you should always avoid interpreting the regression coefficient estimates separately because there is no way to accurately estimate the individual effects of the independent variables. One solution to the problem is to delete the variable with the largest VIF value. The reduced model (i.e., the model with the independent variable with the largest VIF value deleted) is often free of collinearity problems. If you determine that all the independent variables are needed in the model, you can use methods discussed in reference 1.

In the OmniPower sales data (see Section 14.1), the correlation between the two independent variables, price and promotional expenditure, is -0.0968 . Because there are only two independent variables in the model, from Equation (15.8):

$$\begin{aligned} VIF_1 &= VIF_2 = \frac{1}{1 - (-0.0968)^2} \\ &= 1.009 \end{aligned}$$

Thus, you can conclude that you should not be concerned with collinearity for the OmniPower sales data.

In models containing quadratic and interaction terms, collinearity is usually present. The linear and quadratic terms of an independent variable are usually highly correlated with each other, and an interaction term is often correlated with one or both of the independent variables making up the interaction. Thus, you cannot interpret individual regression coefficients separately. You need to interpret the linear and quadratic regression coefficients together in order to understand the nonlinear relationship. Likewise, you need to interpret an interaction regression coefficient in conjunction with the two regression coefficients associated with the variables comprising the interaction. In summary, large *VIFs* in quadratic or interaction models do not necessarily mean that the model is not a good one. They do, however, require you to carefully interpret the regression coefficients.

Problems for Section 15.3

LEARNING THE BASICS

15.14 If the coefficient of determination between two independent variables is 0.20, what is the *VIF*?

15.15 If the coefficient of determination between two independent variables is 0.50, what is the *VIF*?

APPLYING THE CONCEPTS

 **15.16** Refer to Problem 14.4 on page 548. Perform a multiple regression analysis using the data in **WareCost** and determine the *VIF* for each independent variable in the model. Is there reason to suspect the existence of collinearity?

15.17 Refer to Problem 14.5 on page 549. Perform a multiple regression analysis using the data in **VinhoVerde** and determine

the *VIF* for each independent variable in the model. Is there reason to suspect the existence of collinearity?

15.18 Refer to Problem 14.6 on page 549. Perform a multiple regression analysis using the data in **Advertise** and determine the *VIF* for each independent variable in the model. Is there reason to suspect the existence of collinearity?

15.19 Refer to Problem 14.7 on page 549. Perform a multiple regression analysis using the data in **Standby** and determine the *VIF* for each independent variable in the model. Is there reason to suspect the existence of collinearity?

15.20 Refer to Problem 14.8 on page 549. Perform a multiple regression analysis using the data in **GlenCove** and determine the *VIF* for each independent variable in the model. Is there reason to suspect the existence of collinearity?

15.4 Model Building

This chapter and Chapter 14 have introduced you to many different topics in regression analysis, including quadratic terms, dummy variables, and interaction terms. In this section, you learn a structured approach to building the most appropriate regression model. As you will see, successful model building incorporates many of the topics you have studied so far.

To begin, refer to the WSTA-TV scenario introduced on page 596, in which four independent variables (total staff present, remote hours, Dubner hours, and total labor hours) are considered in the business problem that involves developing a regression model to predict standby hours of unionized graphic artists. Data are collected over a period of 26 weeks and organized and stored in **Standby**. Table 15.2 summarizes these data.

TABLE 15.2

Predicting Standby Hours Based on Total Staff Present, Remote Hours, Dubner Hours, and Total Labor Hours

Week	Standby Hours (Y)	Total Staff Present (X_1)	Remote Hours (X_2)	Dubner Hours (X_3)	Total Labor Hours (X_4)
1	245	338	414	323	2,001
2	177	333	598	340	2,030
3	271	358	656	340	2,226
4	211	372	631	352	2,154
5	196	339	528	380	2,078
6	135	289	409	339	2,080
7	195	334	382	331	2,073
8	118	293	399	311	1,758
9	116	325	343	328	1,624
10	147	311	338	353	1,889
11	154	304	353	518	1,988
12	146	312	289	440	2,049
13	115	283	388	276	1,796

(continued)

TABLE 15.2

(continued)

Week	Standby Hours (Y)	Total Staff Present (X ₁)	Remote Hours (X ₂)	Dubner Hours (X ₃)	Total Labor Hours (X ₄)
14	161	307	402	207	1,720
15	274	322	151	287	2,056
16	245	335	228	290	1,890
17	201	350	271	355	2,187
18	183	339	440	300	2,032
19	237	327	475	284	1,856
20	175	328	347	337	2,068
21	152	319	449	279	1,813
22	188	325	336	244	1,808
23	188	322	267	253	1,834
24	197	317	235	272	1,973
25	261	315	164	223	1,839
26	232	331	270	272	1,935

To develop a model to predict the dependent variable, standby hours in the WSTA-TV scenario, you need to be guided by a general problem-solving strategy, or *heuristic*. One heuristic appropriate for building regression models uses the principle of parsimony.

Parsimony guides you to select the regression model with the fewest independent variables that can predict the dependent variable adequately. Regression models with fewer independent variables are easier to interpret, particularly because they are less likely to be affected by collinearity problems (described in Section 15.3).

Developing an appropriate model when many independent variables are under consideration involves complexities that are not present with a model that has only two independent variables. The evaluation of all possible regression models is more computationally complex. And, although you can quantitatively evaluate competing models, there may not be a *uniquely best model* but several *equally appropriate* models.

To begin analyzing the standby-hours data, you compute the variance inflationary factors [see Equation (15.8) on page 608] to measure the amount of collinearity among the independent variables. The values for the four *VIFs* for this model appear in Figure 15.10, along with the results for the model that uses the four independent variables.

FIGURE 15.10

Excel and Minitab regression results for predicting standby hours based on four independent variables (Excel results contain additional worksheets for Durbin-Watson statistic, inset, and VIF)

Standby Hours Analysis					
Regression Statistics					
Multiple R					0.7894
R Square					0.6231
Adjusted R Square					0.5513
Standard Error					31.8350
Observations					26
Durbin-Watson Calculations					
Sum of Squared Difference of Residuals					47241.6126
Sum of Squared Residuals					21282.8217
Durbin-Watson Statistic					2.2197
ANOVA					
	df	SS	MS	F	Significance F
Regression	4	35181.7937	8795.4484	8.6786	0.0003
Residual	21	21282.8217	1013.4677		
Total	25	56464.6154			
	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	-330.8318	110.8954	-2.9833	0.0071	-561.4514
Total Staff	1.2456	0.4121	3.0229	0.0065	0.3887
Remote	-0.1184	0.0543	-2.1798	0.0408	-0.2314
Dubner	-0.2971	0.1179	-2.5189	0.0199	-0.5423
Total Labor	0.1305	0.0593	2.2004	0.0391	0.0072
					0.2539
Variance Inflationary Factor (VIF) Calculations					
	Regression Model				
	Total Staff and all other X	Remote and all other X	Dubner and all other X	Total Labor and all other X	
R Square	0.4143	0.1891	0.3147	0.4998	
VIF	1.7074	1.2333	1.4592	1.9993	

Regression Analysis: Standby versus Total Staff, ...
The regression equation is
Standby = - 330 + 1.25 Total Staff - 0.118 Remote
- 0.297 Dubner + 0.131 Total Labor

Predictor	Coef	SE Coef	T	P	VIF
Constant	-330.8	110.9	-2.98	0.007	
Total Staff	1.2456	0.4121	3.02	0.006	1.707
Remote	-0.1184	0.05432	-2.18	0.041	1.233
Dubner	-0.2971	0.1179	-2.52	0.020	1.459
Total Labor	0.13053	0.05932	2.20	0.039	1.999

S = 31.8350 R-Sq = 62.3% R-Sq(adj) = 55.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	35182	8795	8.68	0.000
Residual Error	21	21283	1013		
Total	25	56465			

Source DF Seq SS
Total Staff 1 20667
Remote 1 6995
Dubner 1 2612
Total Labor 1 4907

Durbin-Watson statistic = 2.21971

Observe that all the *VIF* values in Figure 15.10 are relatively small, ranging from a high of 1.9993 for the total labor hours to a low of 1.2333 for remote hours. Thus, on the basis of the criteria developed by Snee that all *VIF* values should be less than 5.0 (see reference 5), there is little evidence of collinearity among the set of independent variables.

The Stepwise Regression Approach to Model Building

You continue your analysis of the standby-hours data by attempting to determine whether a subset of all independent variables yields an adequate and appropriate model. The first approach described here is **stepwise regression**, which attempts to find the “best” regression model without examining all possible models.

The first step of stepwise regression is to find the best model that uses one independent variable. The next step is to find the best of the remaining independent variables to add to the model selected in the first step. An important feature of the stepwise approach is that an independent variable that has entered into the model at an early stage may subsequently be removed after other independent variables are considered. Thus, in stepwise regression, variables are either added to or deleted from the regression model at each step of the model-building process. The *t* test for the slope (see Section 14.4) or the partial F_{STAT} test statistic (see Section 14.5) is used to determine whether variables are added or deleted. The stepwise procedure terminates with the selection of a best-fitting model when no additional variables can be added to or deleted from the last model evaluated. Figure 15.11 below shows the Excel (PHStat) and Minitab stepwise regression results for the standby-hours data.

FIGURE 15.11

Excel (PHStat) and Minitab stepwise regression results for the standby-hours data

A	B	C	D	E	F	G	H
1 Stepwise Analysis for Standby Hours							
2 Table of Results for General Stepwise							
3							
4 Total Staff entered.							
5							
6	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>		
7	Regression	1	20667.3980	20667.3980	13.8563	0.0011	
8	Residual	24	35797.2174	1491.5507			
9	Total	25	56464.6154				
10							
11	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	
12	Intercept	-272.3816	124.2402	-2.1924	0.0383	-528.8008	-15.9625
13	Total Staff	1.4241	0.3826	3.7224	0.0011	0.6345	2.2136
14							
15							
16	Remote entered.						
17							
18	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>		
19	Regression	2	27662.5429	13831.2714	11.0450	0.0004	
20	Residual	23	28802.0725	1252.2640			
21	Total	25	56464.6154				
22							
23	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	
24	Intercept	-330.6748	116.4802	-2.8389	0.0093	-571.6322	-89.7175
25	Total Staff	1.7649	0.3790	4.6562	0.0001	0.9808	2.5490
26	Remote	-0.1390	0.0588	-2.3635	0.0269	-0.2606	-0.0173
27							
28							
29	No other variables could be entered into the model. Stepwise ends.						

Stepwise Regression: Standby versus Total Staff, Remote, ...						
Alpha-to-Enter: 0.05 Alpha-to-Remove: 0.05						
Response is Standby on 4 predictors, with N = 26						
Step	1		2			
Constant	-272.4		-330.7			
Total Staff	1.42		1.76			
T-Value	3.72		4.66			
P-Value	0.001		0.000			
Remote			-0.139			
T-Value			-2.36			
P-Value			0.027			
S			38.6		35.4	
R-Sq			36.60		48.99	
R-Sq(adj)			33.96		44.56	
Mallows Cp			13.3		8.4	

For this example, a significance level of 0.05 is used to enter a variable into the model or to delete a variable from the model. The first variable entered into the model is total staff, the variable that correlates most highly with the dependent variable standby hours. Because the *p*-value of 0.0011 is less than 0.05, total staff is included in the regression model.

The next step involves selecting a second independent variable for the model. The second variable chosen is one that makes the largest contribution to the model, given that the first variable has been selected. For this model, the second variable is remote hours. Because the *p*-value of 0.0269 for remote hours is less than 0.05, the remote hours variable is included in the regression model.

After the remote hours variable is entered into the model, the stepwise procedure determines whether total staff is still an important contributing variable or whether it can be eliminated from the model. Because the p -value of 0.0001 for total staff is less than 0.05, total staff remains in the regression model.

The next step involves selecting a third independent variable for the model. Because none of the other variables meets the 0.05 criterion for entry into the model, the stepwise procedure terminates with a model that includes total staff present and the number of remote hours.

This stepwise regression approach to model building was originally developed more than four decades ago, when regression computations on computers were time-consuming and costly. Although stepwise regression limited the evaluation of alternative models, the method was deemed a good trade-off between evaluation and cost.

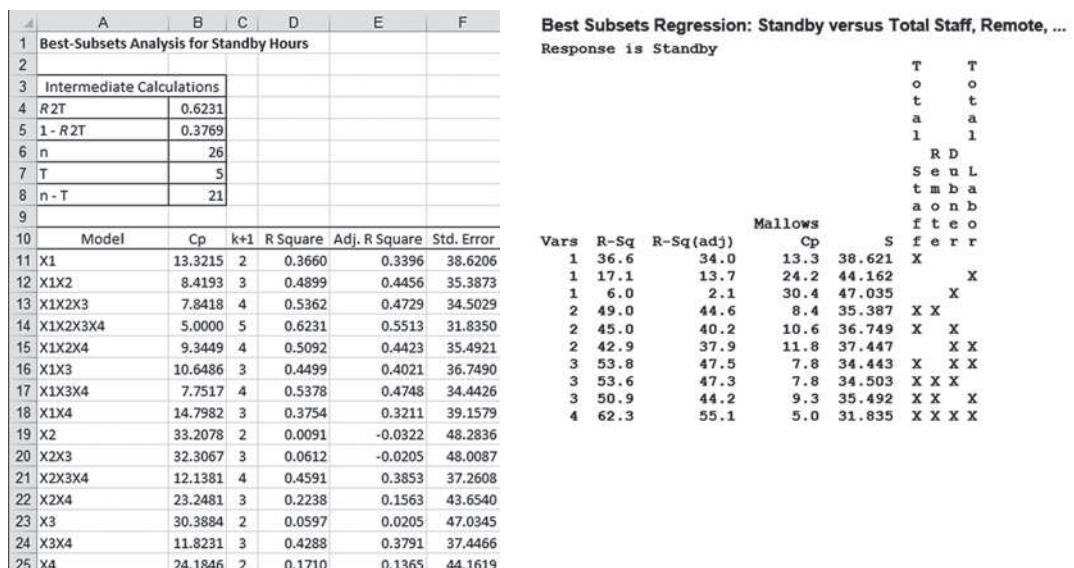
Given the ability of today's computers to perform regression computations at very low cost and high speed, stepwise regression has been superseded to some extent by the best-subsets approach, discussed next, which evaluates a larger set of alternative models. Stepwise regression is not obsolete, however. Today, many businesses use stepwise regression as part of data mining, techniques that allow the extraction of useful knowledge from large repositories of data (see Section 17.2).

The Best-Subsets Approach to Model Building

The **best-subsets approach** evaluates all possible regression models for a given set of independent variables. Figure 15.12 presents best-subsets regression results of all possible regression models for the standby-hours data.

FIGURE 15.12

Excel (using PHStat) and Minitab best-subsets regression results for the standby-hours data



A criterion often used in model building is the adjusted r^2 , which adjusts the r^2 of each model to account for the number of independent variables in the model as well as for the sample size (see Section 14.2). Because model building requires you to compare models with different numbers of independent variables, the adjusted r^2 is more appropriate than r^2 . Referring to Figure 15.12, you see that the adjusted r^2 reaches a maximum value of 0.5513 when all four independent variables plus the intercept term (for a total of five estimated parameters) are included in the model.

A second criterion often used in the evaluation of competing models is the **C_p statistic** developed by Mallows (see reference 1). The C_p statistic, defined in Equation (15.9), measures the differences between a fitted regression model and a *true* model, along with random error.

C_p STATISTIC

$$C_p = \frac{(1 - R_k^2)(n - T)}{1 - R_T^2} - [n - 2(k + 1)] \quad (15.9)$$

where

- k = number of independent variables included in a regression model
- T = total number of parameters (including the intercept) to be estimated in the full regression model
- R_k^2 = coefficient of multiple determination for a regression model that has k independent variables
- R_T^2 = coefficient of multiple determination for a full regression model that contains all T estimated parameters

Using Equation (15.9) to compute C_p for the model containing total staff and remote hours,

$$n = 26 \quad k = 2 \quad T = 4 + 1 = 5 \quad R_k^2 = 0.4899 \quad R_T^2 = 0.6231$$

so that

$$\begin{aligned} C_p &= \frac{(1 - 0.4899)(26 - 5)}{1 - 0.6231} - [26 - 2(2 + 1)] \\ &= 8.4193 \end{aligned}$$

When a regression model with k independent variables contains only random differences from a *true* model, the mean value of C_p is $k + 1$, the number of parameters. Thus, in evaluating many alternative regression models, the goal is to find models whose C_p is close to or less than $k + 1$. In Figure 15.12, you see that only the model with all four independent variables considered contains a C_p value close to or below $k + 1$. Therefore, using the C_p criterion, you should choose that model.

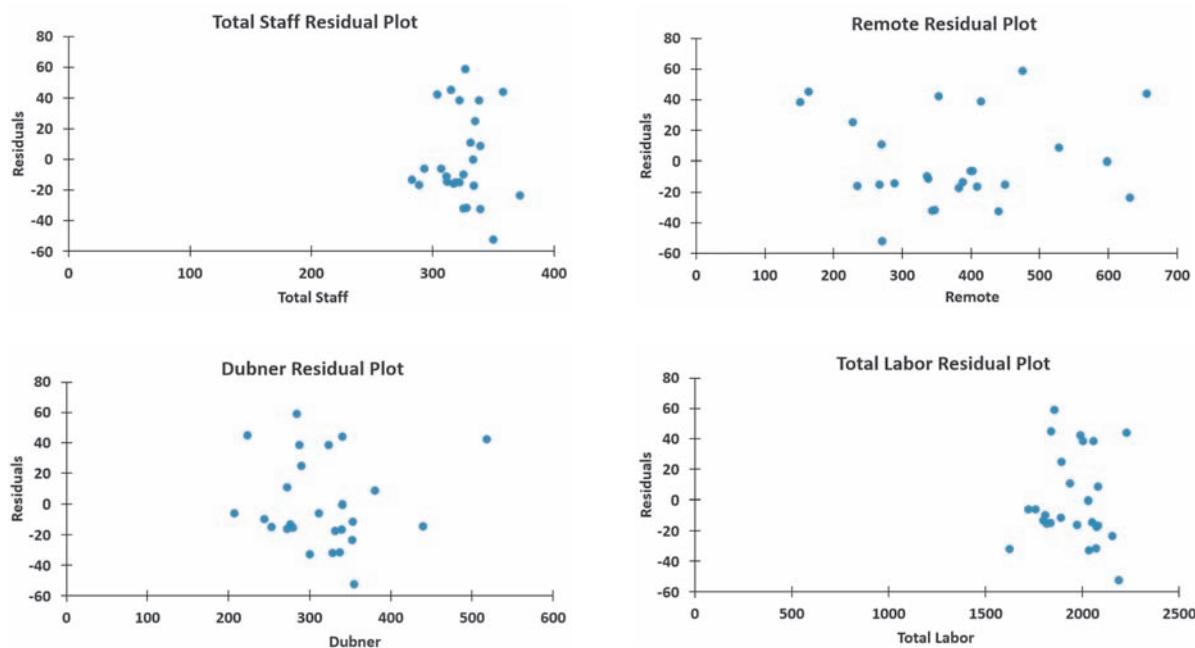
Although it is not the case here, the C_p statistic often provides several alternative models for you to evaluate in greater depth. Moreover, the best model or models using the C_p criterion might differ from the model selected using the adjusted r^2 and/or the model selected using the stepwise procedure. (Note here that the model selected using stepwise regression has a C_p value of 8.4193, which is substantially above the suggested criterion of $k + 1 = 3$ for that model.) Remember that there may not be a uniquely best model, but there may be several equally appropriate models. Final model selection often involves using subjective criteria, such as parsimony, interpretability, and departure from model assumptions (as evaluated by residual analysis).

When you have finished selecting the independent variables to include in the model, you need to perform a residual analysis to evaluate the regression assumptions, and because the data were collected in time order, you also need to compute the Durbin-Watson statistic to determine whether there is autocorrelation in the residuals (see Section 13.6). From Figure 15.10 on page 610, you see that the Durbin-Watson statistic, D , is 2.2197. Because D is greater than 2.0, there is no indication of positive correlation in the residuals. Figure 15.13 on page 614 presents the plots used in the residual analysis.

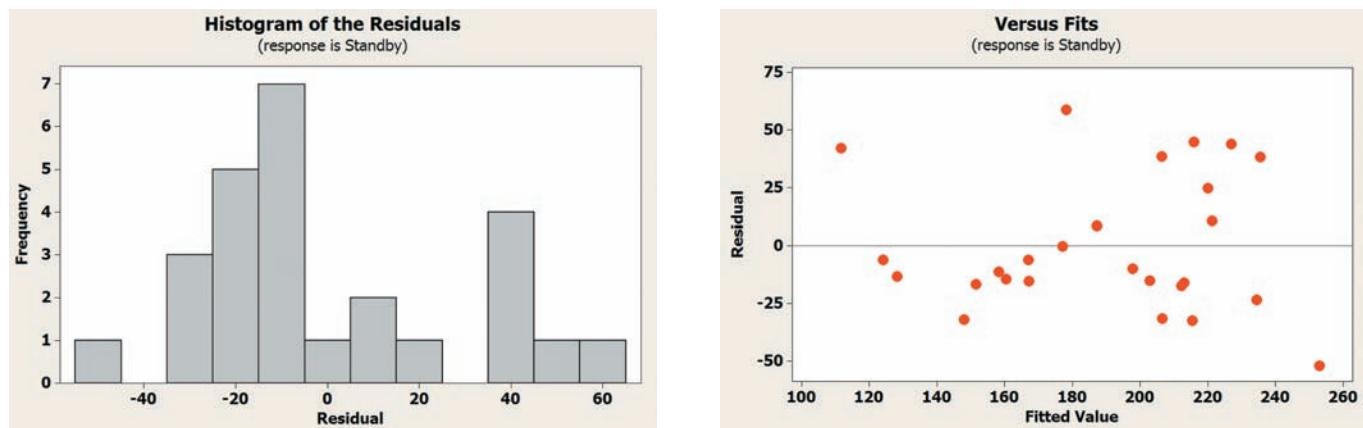
None of the residual plots versus the total staff, the remote hours, the Dubner hours, and the total labor hours reveal apparent patterns. In addition, a histogram of the residuals (shown in Figure 15.14 on page 614) indicates only moderate departure from normality, and a plot of the residuals versus the predicted values of Y (also shown in Figure 15.14) does not show evidence of unequal variance.

FIGURE 15.13

Residual plots for the standby-hours data

**FIGURE 15.14**

Histogram of the residuals and scatter plot of the residuals versus the predicted values of Y


Student Tip

You can use Table E.3 to find the critical t value with $n - k - 2 = 26 - 4 - 2 = 20$ degrees of freedom.

Because the residual analysis appears to confirm the aptness of the model, you can now use various influence measures discussed in Section 14.8 to determine whether any of the values unduly influence the regression equation. Figure 15.15 on page 615 presents the h_i , t_i , and Cook's D_i statistics for the fitted model and highlights certain observations for further analysis.

Using the decision rule suggested by Hoaglin and Welsch, you flag any observation that has a h_i value greater than $2(k + 1)/n = 2(4 + 1)/26 = 0.3846$. From Figure 15.15, you see that observations 6 ($h_6 = 0.4049$), 9 ($h_9 = 0.4537$), and 11 ($h_{11} = 0.5217$) have h_i values that are greater than 0.3846 and therefore are candidates for deletion from the analysis.

You next apply the Studentized deleted residual measure t_i to the influence analysis. Using the decision rule suggested by Hoaglin and Welsch, any observation with t_i value greater than 1.7247 or less than -1.7247 should be flagged. From Figure 15.15, you see that $t_3 = 1.741$, $t_{11} = 2.0673$, $t_{17} = -2.0072$, and $t_{19} = 2.1029$, which suggests that observations 3, 11, 17, and 19 may have an adverse effect on the model. You note that observation 11 is also a candidate for deletion according to the h_i criterion.

FIGURE 15.15

Minitab influence statistics for the standby hours data

+	C1	C2	C3	C4	C5	C6	C7	C8
	Standby	Total Staff	Remote	Dubner	Total Labor	TRES1	H1	COOK1
1	245	338	414	323	2001	1.26648	0.057851	0.019147
2	177	333	598	340	2030	-0.00450	0.159009	0.000001
3	271	358	656	340	2226	1.74109	0.308626	0.246769
4	211	372	631	352	2154	-0.88635	0.317663	0.073903
5	196	339	528	380	2078	0.28517	0.117963	0.002275
6	135	289	409	339	2080	-0.66415	0.404904	0.061665
7	195	334	382	331	2073	-0.55135	0.066692	0.004493
8	118	293	399	311	1758	-0.20251	0.177819	0.001859
9	116	325	343	328	1624	-1.38665	0.453697	0.305928
10	147	311	338	353	1889	-0.36082	0.080160	0.002367
11	154	304	353	518	1988	2.06732	0.521708	0.806606
12	146	312	289	440	2049	-0.50740	0.239542	0.016814
13	115	283	388	276	1796	-0.47510	0.267786	0.017142
14	161	307	402	207	1720	-0.20841	0.219149	0.002554
15	274	322	151	287	2056	1.42189	0.241543	0.122799
16	245	335	228	290	1890	0.84801	0.155157	0.026771
17	201	350	271	355	2187	-2.00716	0.240144	0.222548
18	183	339	440	300	2032	-1.06250	0.073309	0.017752
19	237	327	475	284	1856	2.10290	0.101265	0.085690
20	175	328	347	337	2068	-1.02770	0.071640	0.016257
21	152	319	449	279	1813	-0.49356	0.105621	0.005969
22	188	325	336	244	1808	-0.31659	0.107193	0.002515
23	188	322	267	253	1834	-0.48404	0.100514	0.005434
24	197	317	235	272	1973	-0.52597	0.123908	0.008105
25	261	315	164	223	1839	1.63790	0.193025	0.118818
26	232	331	270	272	1935	0.34618	0.094110	0.002599

You then apply a third criterion, Cook's D_i statistic, based on h_i and the standardized residual. For the model, in which $k = 4$ and $n = 26$, using the decision rule suggested by Cook and Weisberg, you flag any $D_i > 0.899$, the critical value for the F statistic having 5 and 21 degrees of freedom at the 0.50 level of significance (see Table 14.15 on page 580). From Figure 15.15, none of the D_i values exceed 0.899, including the D_i for observation 11 (0.807). You note that according to Cook's D_i statistic, no observations are candidates for deletion.

With this influence analysis, you conclude that you have no clear basis for removing any of the observations from the analysis. Therefore, using the Figure 15.10 regression model results, you state the regression equation as

$$\hat{Y}_i = -330.8318 + 1.2456X_{1i} - 0.1184X_{2i} - 0.2971X_{3i} + 0.1305X_{4i}$$

Example 15.4 presents a situation in which there are several alternative models in which the C_p statistic is close to or less than $k + 1$.

EXAMPLE 15.4

Choosing Among Alternative Regression Models

TABLE 15.3
Partial Results from Best-Subsets Regression

Table 15.3 shows results from a best-subsets regression analysis of a regression model with seven independent variables. Determine which regression model you would choose as the *best* model.

Number of Variables	r^2	Adjusted r^2	C_p	Variables Included
1	0.121	0.119	113.9	X_4
1	0.093	0.090	130.4	X_1
1	0.083	0.080	136.2	X_3
2	0.214	0.210	62.1	X_3, X_4
2	0.191	0.186	75.6	X_1, X_3
2	0.181	0.177	81.0	X_1, X_4
3	0.285	0.280	22.6	X_1, X_3, X_4
3	0.268	0.263	32.4	X_3, X_4, X_5
3	0.240	0.234	49.0	X_2, X_3, X_4
4	0.308	0.301	11.3	X_1, X_2, X_3, X_4
4	0.304	0.297	14.0	X_1, X_3, X_4, X_6
4	0.296	0.289	18.3	X_1, X_3, X_4, X_5
5	0.317	0.308	8.2	X_1, X_2, X_3, X_4, X_5
5	0.315	0.306	9.6	X_1, X_2, X_3, X_4, X_6
5	0.313	0.304	10.7	X_1, X_3, X_4, X_5, X_6
6	0.323	0.313	6.8	$X_1, X_2, X_3, X_4, X_5, X_6$
6	0.319	0.309	9.0	$X_1, X_2, X_3, X_4, X_5, X_7$
6	0.317	0.306	10.4	$X_1, X_2, X_3, X_4, X_6, X_7$
7	0.324	0.312	8.0	$X_1, X_2, X_3, X_4, X_5, X_6, X_7$

(continued)

SOLUTION From Table 15.3, you need to determine which models have C_p values that are less than or close to $k + 1$. Two models meet this criterion. The model with six independent variables ($X_1, X_2, X_3, X_4, X_5, X_6$) has a C_p value of 6.8, which is less than $k + 1 = 6 + 1 = 7$, and the full model with seven independent variables ($X_1, X_2, X_3, X_4, X_5, X_6, X_7$) has a C_p value of 8.0. One way you can choose among the two models is to select the model with the largest adjusted r^2 —that is, the model with six independent variables. Another way to select a final model is to determine whether the models contain a subset of variables that are common. Then you test whether the contribution of the additional variables is significant. In this case, because the models differ only by the inclusion of variable X_7 in the full model, you test whether variable X_7 makes a significant contribution to the regression model, given that the variables X_1, X_2, X_3, X_4, X_5 , and X_6 are already included in the model. If the contribution is statistically significant, then you should include variable X_7 in the regression model. If variable X_7 does not make a statistically significant contribution, you should not include it in the model.

Model Validation

The final step in the model-building process is to validate the selected regression model. This step involves checking the model against data that were not part of the sample analyzed. The following are several ways of validating a regression model:

- Collect new data and compare the results.
- Compare the results of the regression model to previous results.
- If the data set is large, split the data into two parts and cross-validate the results.

Perhaps the best way of validating a regression model is by collecting new data. If the results with new data are consistent with the selected regression model, you have strong reason to believe that the fitted regression model is applicable in a wide set of circumstances.

If it is not possible to collect new data, you can use one of the two other approaches. In one approach, you compare your regression coefficients and predictions to previous results. If the data set is large, you can use **cross-validation**. First, you split the data into two parts. Then you use the first part of the data to develop the regression model. You then use the second part of the data to evaluate the predictive ability of the regression model.

Steps for successful model building are summarized below and in Figure 15.16.

Steps for Successful Model Building

The following steps summarize the steps for successful model building discussed in Section 15.4.

Step 1 Choose the independent variables to be considered for the multiple regression model.

Step 2 Develop a regression model that includes all of the independent variables.

Step 3 Compute the VIF for each independent variable.

If none of the independent variables has a $VIF > 5$, continue with step 4.

If only one of the independent variables has a $VIF > 5$, eliminate that independent variable and continue with step 4.

If more than one of the independent variables has a $VIF > 5$, eliminate the independent variable that has the highest VIF from the regression model and repeat step 2 with the remaining independent variables.

Step 4 Perform a best-subsets regression with the remaining independent variables and compute the C_p statistic and/or the adjusted r^2 for each model.

Step 5 List all models that have C_p close to or less than $k + 1$ and/or a high adjusted r^2 .

Step 6 From the models listed in step 4, choose a best model.

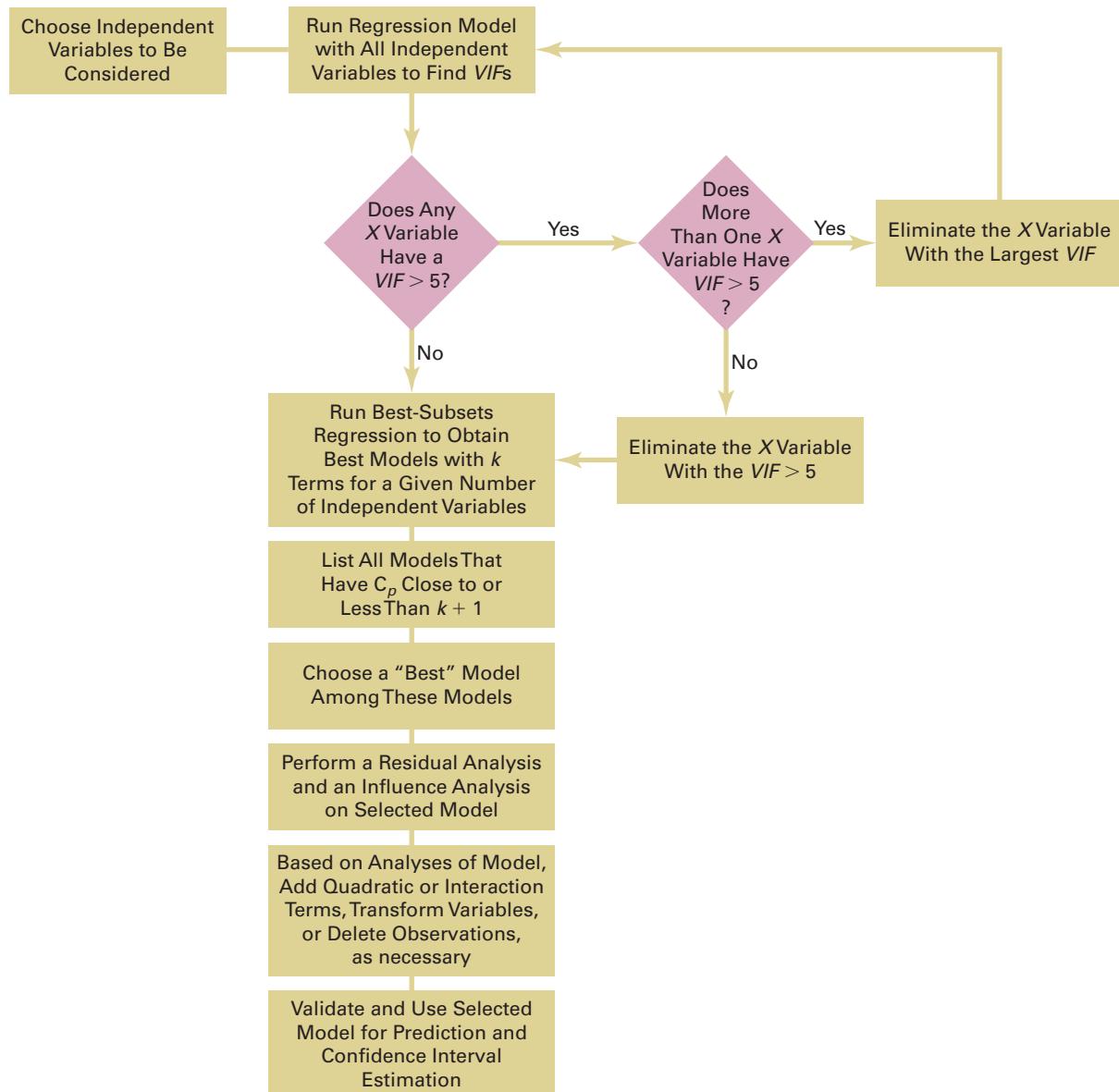
Step 7 Perform a complete analysis of the model chosen, including a residual analysis and influence analysis.

Step 8 Review the results of the residual analysis and influence analysis. If necessary, add quadratic and/or interaction terms, transform variables, and delete individual observations, as necessary, and repeat steps 2 through 6. If no changes are necessary, continue with step 8.

Step 9 Validate the regression model. If validated, use the regression model for prediction and inference.

FIGURE 15.16

Roadmap for model building



Problems for Section 15.4

LEARNING THE BASICS

15.21 You are considering four independent variables for inclusion in a regression model. You select a sample of $n = 30$, with the following results:

1. The model that includes independent variables A and B has a C_p value equal to 4.6.
2. The model that includes independent variables A and C has a C_p value equal to 2.4.
3. The model that includes independent variables A , B , and C has a C_p value equal to 2.7.
- a. Which models meet the criterion for further consideration? Explain.

- b. How would you compare the model that contains independent variables A , B , and C to the model that contains independent variables A and B ? Explain.

15.22 You are considering six independent variables for inclusion in a regression model. You select a sample of $n = 40$, with the following results:

$$k = 2 \quad T = 6 + 1 = 7 \quad R_k^2 = 0.274 \quad R_T^2 = 0.653$$

- a. Compute the C_p value for this two-independent-variable model.
- b. Based on your answer to (a), does this model meet the criterion for further consideration as the best model? Explain.

APPLYING THE CONCEPTS

15.23 The file **FTMBA** contains a sample of 2012 top-ranked full-time MBA programs. Variables included are mean starting salary upon graduation (\$), percent of students with job offers within three months of graduation, program cost (\$), and total number of students per program. (Data extracted from buswk.co/25d1ZC.) Develop the most appropriate multiple regression model to predict the mean starting salary upon graduation. Be sure to include a thorough residual and influence analysis. In addition, provide a detailed explanation of the results, including a comparison of the most appropriate multiple regression model to the best simple linear regression model.

 **15.24** You need to develop a model to predict the assessed value of houses in Silver Spring, Maryland, based on the size of the house, whether it has a fireplace, the number of bedrooms, and the number of bathrooms. A sample of 30 houses is selected and the results are stored in **SilverSpring**.

Develop the most appropriate multiple regression model to predict assessed value. Be sure to perform a thorough residual and influence analysis. In addition, provide a detailed explanation of the results.

15.25 *Accounting Today* identified top public accounting firms in ten geographic regions across the U.S. The file **AccountingPartners6** contains data for public accounting firms in the Southeast, Gulf Coast, and Capital Regions. The variables are: revenue (\$M), number of partners in the firm, number of professionals in the firm, proportion of business dedicated to management advisory services (MAS%), whether the firm is located in the Southeast Region (0 = no, 1 = yes), and whether the firm is located in the Gulf Coast Region (0 = no, 1 = yes). (Data extracted from bit.ly/176TPfR.)

Develop the most appropriate multiple regression model to predict firm revenue. Be sure to perform a thorough residual and influence analysis. In addition, provide a detailed explanation of the results.

15.5 Pitfalls in Multiple Regression and Ethical Issues

Pitfalls in Multiple Regression

Model building is an art as well as a science. Different individuals may not always agree on the best multiple regression model. To construct a good regression model, you should use the process described on page 616. In doing so, you must avoid certain pitfalls that can interfere with the development of a useful model. Section 13.9 discussed pitfalls in simple linear regression and strategies for avoiding them. Now that you have studied a variety of multiple regression models, you need to take some additional precautions. To avoid pitfalls in multiple regression, you also need to

- Interpret the regression coefficient for a particular independent variable from a perspective in which the values of all other independent variables are held constant.
- Evaluate residual plots for each independent variable.
- Evaluate interaction and quadratic terms.
- Compute the *VIF* for each independent variable before determining which independent variables to include in the model.
- Examine several alternative models, using best-subsets regression.
- Use influence analysis to determine whether to remove any observations from the analysis.
- Use logistic regression instead of least squares regression when the dependent variable is categorical.
- Validate the model before implementing it.

Ethical Issues

Ethical issues arise when a user who wants to make predictions manipulates the development process of the multiple regression model. The key here is intent. In addition to the situations discussed in Section 13.9, unethical behavior occurs when someone uses multiple regression analysis and *willfully fails* to remove from consideration independent variables that exhibit a high collinearity with other independent variables or *willfully fails* to use methods other than least-squares regression when the assumptions necessary for least-squares regression are seriously violated.

USING STATISTICS

Valuing Parsimony at WSTA-TV, Revisited

In the Using Statistics scenario, you were the WSTA-TV broadcast operations manager, who sought to reduce labor expenses. You needed to determine which variables have an effect on standby hours, the time during which graphic artists employed by the station are idle but getting paid. You collected data concerning standby hours and the total number of staff present, remote hours, Dubner hours, and total labor hours over a period of 26 weeks.

You performed a multiple regression analysis on the data. The coefficient of multiple determination indicated that 62.31% of the variation in standby hours can be explained by variation in the number of graphic artists present and the number of remote hours, Dubner hours, and total labor hours. The model indicated that standby hours are estimated to increase by 1.2456 hours for each additional staff member present, holding constant the other independent

variables; to decrease by 0.1184 hour for each additional remote hour, holding constant the other independent variables; to decrease by 0.2971 hour for each additional Dubner hour, holding constant the other independent variables; and to increase by 0.1305 hour for each additional labor hour, holding constant the other independent variables. Each of the four independent variables had a significant effect on standby hours, holding constant the other independent variables. This regression model enables you to predict standby hours based on the total number of graphic artists present, remote hours, Dubner hours, and total labor hours. Any predictions developed by the model can then be carefully monitored, new data can be collected, and other variables may possibly be considered.



Alamy

SUMMARY

This chapter discussed quadratic regression models, transformations, collinearity, and model building. The

Figure 15.17 roadmap on page 620 summarizes and interrelates these topics. (Chapter 15 references appear on page 621.)

KEY EQUATIONS

Quadratic Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i \quad (15.1)$$

Quadratic Regression Equation

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{1i}^2 \quad (15.2)$$

Regression Model with a Square-Root Transformation

$$Y_i = \beta_0 + \beta_1 \sqrt{X_{1i}} + \varepsilon_i \quad (15.3)$$

Original Multiplicative Model

$$Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} \varepsilon_i \quad (15.4)$$

Transformed Multiplicative Model

$$\begin{aligned} \log Y_i &= \log(\beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} \varepsilon_i) \\ &= \log \beta_0 + \log(X_{1i}^{\beta_1}) + \log(X_{2i}^{\beta_2}) + \log \varepsilon_i \\ &= \log \beta_0 + \beta_1 \log X_{1i} + \beta_2 \log X_{2i} + \log \varepsilon_i \end{aligned} \quad (15.5)$$

Original Exponential Model

$$Y_i = e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}} \varepsilon_i \quad (15.6)$$

Transformed Exponential Model

$$\begin{aligned} \ln Y_i &= \ln(e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}} \varepsilon_i) \\ &= \ln(e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}}) + \ln \varepsilon_i \\ &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ln \varepsilon_i \end{aligned} \quad (15.7)$$

Variance Inflationary Factor

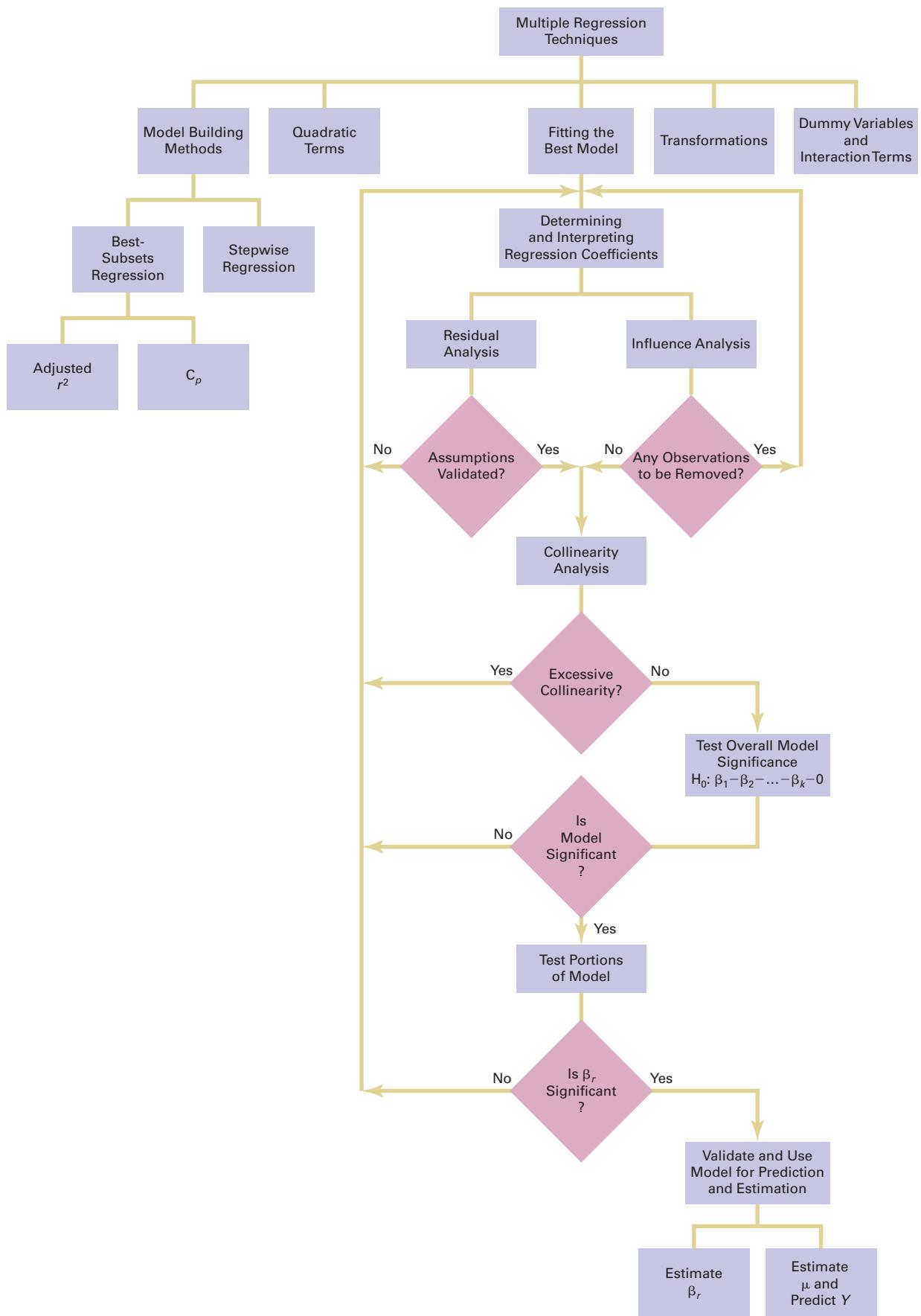
$$VIF_j = \frac{1}{1 - R_j^2} \quad (15.8)$$

C_p Statistic

$$C_p = \frac{(1 - R_k^2)(n - T)}{1 - R_T^2} - [n - 2(k + 1)] \quad (15.9)$$

FIGURE 15.17

Roadmap for multiple regression



REFERENCES

1. Kutner, M., C. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*, 5th ed. New York: McGraw-Hill/Irwin, 2005.
2. Marquardt, D. W. W. "Comment on Smith and Campbell (1980)." *Journal of the American Statistical Association*, 75 (1980): 87–91.
3. Microsoft Excel 2013. Redmond, WA: Microsoft Corp., 2012.
4. Minitab Release 16. State College, PA: Minitab, Inc., 2010.
5. Snee, R. D. "Some Aspects of Nonorthogonal Data Analysis, Part I. Developing Prediction Equations." *Journal of Quality Technology* 5 (1973): 67–79.

KEY TERMS

best-subsets approach 612
 C_p statistic 612
 collinearity 608
 cross-validation 616

logarithmic transformation 605
 parsimony 610
 quadratic regression model 597
 quadratic term 597

square-root transformation 605
 stepwise regression 611
 variance inflationary factor (*VIF*) 608

CHECKING YOUR UNDERSTANDING

15.26 How can you evaluate whether collinearity exists in a multiple regression model?

15.27 What is the difference between stepwise regression and best-subsets regression?

15.28 How do you choose among models according to the C_p statistic in best-subsets regression?

CHAPTER REVIEW PROBLEMS

15.29 A specialist in baseball analytics has expanded his analysis, presented in Problem 14.81 on page 586, of which variables are important in predicting a team's wins in a given baseball season. He has collected data in **BB2012** related to wins, ERA, saves, runs scored, hits allowed, walks allowed, and errors for the 2012 season.

- Develop the most appropriate multiple regression model to predict a team's wins. Be sure to include a thorough residual and influence analysis. In addition, provide a detailed explanation of the results.
- Develop the most appropriate multiple regression model to predict a team's ERA on the basis of hits allowed, walks allowed, errors, and saves. Be sure to include a thorough residual and influence analysis. In addition, provide a detailed explanation of the results.

15.30 In the production of printed circuit boards, errors in the alignment of electrical connections are a source of scrap. The file **RegistrationError** contains the registration error, the temperature, the pressure, and the cost of the material (low versus high) used in the production of circuit boards. (Data extracted from C. Nachtsheim and B. Jones, "A Powerful Analytical Tool," *Six Sigma Forum Magazine*, August 2003, pp. 30–33.) Develop the most appropriate multiple regression model to predict registration error.

15.31 Hemlock Farms is a community located in the Pocono Mountains area of eastern Pennsylvania. The file **HemlockFarms** contains information on homes that were recently for sale. The variables included were

List Price—Asking price of the house

Hot Tub—Whether the house has a hot tub, with 0 = No and 1 = Yes

Lake View—Whether the house has a lake view, with 0 = No and 1 = Yes

Bathrooms—Number of bathrooms

Bedrooms—Number of bedrooms

Loft/Den—Whether the house has a loft or den, with 0 = No and 1 = Yes

Finished basement—Whether the house has a finished basement, with 0 = No and 1 = Yes

Acres—Number of acres for the property

Develop the most appropriate multiple regression model to predict the asking price. Be sure to perform a thorough residual and influence analysis. In addition, provide a detailed explanation of your results.

15.32 Nassau County is located approximately 25 miles east of New York City. The file **GlenCove** contains a sample of 30 single-family homes located in Glen Cove. Variables included are the fair market value, land area of the property (acres), interior size of the house (square feet), age (years), number of rooms, number of bathrooms, and number of cars that can be parked in the garage.

- Develop the most appropriate multiple regression model to predict fair market value.
- Compare the results in (a) with those of Problems 15.33 (a) and 15.34 (a).

15.33 Data similar to those in Problem 15.32 are available for homes located in Roslyn (approximately 8 miles from Glen Cove) and are stored in **Roslyn**.

- Perform an analysis similar to that of Problem 15.32.
- Compare the results in (a) with those of Problems 15.32 (a) and 15.34 (a).

15.34 Data similar to Problem 15.32 are available for homes located in Freeport (located approximately 20 miles from Roslyn) and are stored in **Freeport**.

- Perform an analysis similar to that of Problem 15.32.
- Compare the results in (a) with those of Problems 15.32 (a) and 15.33 (a).

15.35 You are a real estate broker who wants to compare property values in Glen Cove and Roslyn (which are located approximately 8 miles apart). Use the data in **GCRoslyn**. Make sure to include the dummy variable for location (Glen Cove or Roslyn) in the regression model.

- Develop the most appropriate multiple regression model to predict fair market value.
- What conclusions can you reach concerning the differences in fair market value between Glen Cove and Roslyn?

15.36 You are a real estate broker who wants to compare property values in Glen Cove, Freeport, and Roslyn. Use the data in **GCFreeRoslyn**.

- Develop the most appropriate multiple regression model to predict fair market value.
- What conclusions can you reach concerning the differences in fair market value between Glen Cove, Freeport, and Roslyn?

15.37 Financial analysts engage in business valuation to determine a company's value. A standard approach uses the multiple of earnings method: You multiply a company's profits by a certain value (average or median) to arrive at a final value. More recently, regression analysis has been demonstrated to consistently deliver more accurate predictions. A valuator has been given the assignment of valuing a drug company. He obtained financial data on 72 drug companies (Industry Group Standard Industrial Classification [SIC] 3 code 283), which included pharmaceutical preparation firms (SIC 4 code 2834), in vitro and in vivo diagnostic substances firms (SIC 4 code 2835), and biological products firms (SIC 4 2836). The file **BusinessValuation2** contains the following variables:

- COMPANY—Drug company name
- TS—Ticker symbol
- SIC 3—Standard Industrial Classification 3 code (industry group identifier)
- SIC 4—Standard Industrial Classification 4 code (industry identifier)
- PB fy—Price-to-book value ratio (fiscal year ending)
- PE fy—Price-to-earnings ratio (fiscal year ending)
- NL Assets—Natural log of assets (as a measure of size)
- ROE—Return on equity
- SGROWTH—Growth (GS5)
- DEBT/EBITDA—Ratio of debt to earnings before interest, taxes, depreciation, and amortization
- D2834—Dummy variable indicator of SIC 4 code 2834 (1 if 2834, 0 if not)

D2835—Dummy variable indicator of SIC 4 code 2835 (1 if 2835, 0 if not)

Develop the most appropriate multiple regression model to predict the price-to-book value ratio. Be sure to perform a thorough residual and influence analysis. In addition, provide a detailed explanation of your results.

15.38 A recent article (J. Conklin, "It's a Marathon, Not a Sprint," *Quality Progress*, June 2009, pp. 46–49) discussed a metal deposition process in which a piece of metal is placed in an acid bath and an alloy is layered on top of it. The key quality characteristic is the thickness of the alloy layer. The file **Thickness** contains the following variables:

- Thickness—Thickness of the alloy layer
- Catalyst—Catalyst concentration in the acid bath
- pH—pH level of the acid bath
- Pressure—Pressure in the tank holding the acid bath
- Temp—Temperature in the tank holding the acid bath
- Voltage—Voltage applied to the tank holding the acid bath

Develop the most appropriate multiple regression model to predict the thickness of the alloy layer. Be sure to perform a thorough residual and influence analysis. The article suggests that there is a significant interaction between the pressure and the temperature in the tank. Do you agree?

15.39 A molding machine that contains different cavities is used in producing plastic parts. The product characteristics of interest are the product length (in.) and weight (g). The mold cavities were filled with raw material powder and then vibrated during the experiment. The factors that were varied were the vibration time (seconds), the vibration pressure (psi), the vibration amplitude (%), the raw material density (g/mL), and the quantity of raw material (scoops). The experiment was conducted in two different cavities on the molding machine. The data are stored in **Molding**. (Data extracted from M. Lopez and M. McShane-Vaughn, "Maximizing Product, Minimizing Costs," *SixSigma Forum Magazine*, February 2008, pp. 18–23.)

- Develop the most appropriate multiple regression model to predict the product length in cavity 1. Be sure to perform a thorough residual and influence analysis. In addition, provide a detailed explanation of your results.
- Repeat (a) for cavity 2.
- Compare the results for length in the two cavities.
- Develop the most appropriate multiple regression model to predict the product weight in cavity 1. Be sure to perform a thorough residual and influence analysis. In addition, provide a detailed explanation of your results.
- Repeat (d) for cavity 2.
- Compare the results for weight in the two cavities.

REPORT WRITING EXERCISE

15.40 In Problems 15.32–15.36 you developed multiple regression models to predict the fair market value of houses in Glen Cove, Roslyn, and Freeport. Now write a report based on the models you developed. Append all appropriate charts and statistical information to your report.

CASES FOR CHAPTER 15

The Mountain States Potato Company

Mountain States Potato Company sells a by-product of its potato-processing operation, called a filter cake, to area feedlots as cattle feed. The business problem faced by the feedlot owners is that the cattle are not gaining weight as quickly as they once were. The feedlot owners believe that the root cause of the problem is that the percentage of solids in the filter cake is too low.

Historically, the percentage of solids in the filter cakes ran slightly above 12%. Lately, however, the solids are running in the 11% range. What is actually affecting the solids is a mystery, but something has to be done quickly. Individuals involved in the process were asked to identify variables that might affect the percentage of solids. This review turned up the six variables (in addition to the percentage of solids) listed in the right column. Data collected by monitoring the process several times daily for 20 days are stored in **Potato**.

1. Thoroughly analyze the data and develop a regression model to predict the percentage of solids.
2. Write an executive summary concerning your findings to the president of the Mountain States Potato Company. Include specific recommendations on how to get the percentage of solids back above 12%.

Variable	Comments
SOLIDS	Percentage of solids in the filter cake.
PH	Acidity. This measure of acidity indicates bacterial action in the clarifier and is controlled by the amount of downtime in the system. As bacterial action progresses, organic acids are produced that can be measured using pH.
LOWER	Pressure of the vacuum line below the fluid line on the rotating drum.
UPPER	Pressure of the vacuum line above the fluid line on the rotating drum.
THICK	Filter cake thickness, measured on the drum.
VARIDRIV	Setting used to control the drum speed. May differ from DRUMSPD due to mechanical inefficiencies.
DRUMSPD	Speed at which the drum is rotating when collecting the filter cake. Measured with a stopwatch.

Sure Value Convenience Stores

You work in the corporate office for a nationwide convenience store franchise that operates nearly 10,000 stores. The per-store daily customer count (i.e., the mean number of customers in a store in one day) has been steady, at 900, for some time. To increase the customer count, the chain is considering cutting prices for coffee beverages. The question to be determined is how much prices should be cut to increase the daily customer count without reducing the gross margin on coffee sales too much. You decide to carry out an experiment in a sample of 24 stores where customer counts have been running almost exactly at the national average of 900. In six of the stores, the price of a small coffee will now be \$0.59, in six stores the price of a small coffee will now be \$0.69, in six stores, the price of a small coffee will now be \$0.79, and in six stores, the price of a small coffee will now be \$0.89. After four weeks at the new prices, the daily customer count in the stores is determined and is stored in **CoffeeSales2**.

- a. Construct a scatter plot for price and sales.
- b. Fit a quadratic regression model and state the quadratic regression equation.
- c. Predict the mean weekly sales for a small coffee priced at 79 cents.
- d. Perform a residual analysis on the results and determine whether the regression model is valid.
- e. At the 0.05 level of significance, is there a significant quadratic relationship between weekly sales and price?
- f. At the 0.05 level of significance, determine whether the quadratic model is a better fit than the linear model.
- g. Interpret the meaning of the coefficient of multiple determination.
- h. Compute the adjusted r^2 .
- i. What price do you recommend the small coffee should be sold for?

Digital Case

Apply your knowledge of multiple regression model building in this Digital Case, which extends the Chapter 14 OmniPower Bars Using Statistics scenario.

Still concerned about ensuring a successful test marketing of its OmniPower bars, the marketing department of OmniFoods has contacted Connect2Coupons (C2C), another merchandising consultancy. C2C suggests that earlier analysis done by In-Store Placements Group (ISPG) was faulty because it did not use the correct type of data. C2C claims that its Internet-based viral marketing will have an even greater effect on OmniPower energy bar sales, as new data from the same 34-store sample will show. In response, ISPG says its earlier claims are valid and has reported to the OmniFoods marketing

department that it can discern no simple relationship between C2C's viral marketing and increased OmniPower sales.

Open **OmniPowerForum15.pdf** to review all the claims made in a private online forum and chat hosted on the OmniFoods corporate website. Then answer the following:

1. Which of the claims are true? False? True but misleading? Support your answer by performing an appropriate statistical analysis.
2. If the grocery store chain allowed OmniFoods to use an unlimited number of sales techniques, which techniques should it use? Explain.
3. If the grocery store chain allowed OmniFoods to use only one sales technique, which technique should it use? Explain.

The Craybill Instrumentation Company Case

The Craybill Instrumentation Company produces highly technical industrial instrumentation devices. The human resources (HR) director has the business objective of improving recruiting decisions concerning sales managers. The company has 45 sales regions, each headed by a sales manager. Many of the sales managers have degrees in electrical engineering, and due to the technical nature of the product line, several company officials believe that only applicants with degrees in electrical engineering should be considered.

At the time of their application, candidates are asked to take the Strong-Campbell Interest Inventory Test and the Wonderlic Personnel Test. Due to the time and money involved with the testing, some discussion has taken place about dropping one or both of the tests. To start, the HR director gathered information on each of the 45 current sales managers, including years of selling experience, electrical engineering background, and the scores from both the Wonderlic and Strong-Campbell tests. The HR director has decided to use regression modeling to predict a dependent variable of "sales index" score, which is the ratio of the regions' actual sales divided by the target sales. The target values are constructed each year by upper management, in consultation with the sales managers, and are based on past performance and market potential within each region. The file **Managers** contains information on the 45 current sales managers. The following variables are included:

Sales—Ratio of yearly sales divided by the target sales value for that region; the target values were mutually agreed-upon "realistic expectations"

Wonder—Score from the Wonderlic Personnel Test; the higher the score, the higher the applicant's perceived ability to manage

SC—Score on the Strong-Campbell Interest Inventory Test; the higher the score, the higher the applicant's perceived interest in sales

Experience—Number of years of selling experience prior to becoming a sales manager

Engineer—Dummy variable that equals 1 if the sales manager has a degree in electrical engineering and 0 otherwise

- a. Develop the most appropriate regression model to predict sales.
- b. Do you think that the company should continue administering both the Wonderlic and Strong-Campbell tests? Explain.
- c. Do the data support the argument that electrical engineers outperform the other sales managers? Would you support the idea to hire only electrical engineers? Explain.
- d. How important is prior selling experience in this case? Explain.
- e. Discuss in detail how the HR director should incorporate the regression model you developed into the recruiting process.

More Descriptive Choices Follow-Up

Follow-up the Using Statistics scenario "More Descriptive Choices, Revisited" on page 138, by developing regression models to predict the one-year return, the three-year return, the five-year return, and the ten-year return based on the assets, turnover ratio, expense ratio, beta, standard deviation,

type of fund (growth versus value), and risk (stored in **Retirement Funds**). (For this analysis, combine low and average risk into the new category "not high.") Be sure to perform a thorough residual and influence analysis. Provide a summary report that explains your results in detail.

CHAPTER 15 EXCEL GUIDE

EG15.1 The QUADRATIC REGRESSION MODEL

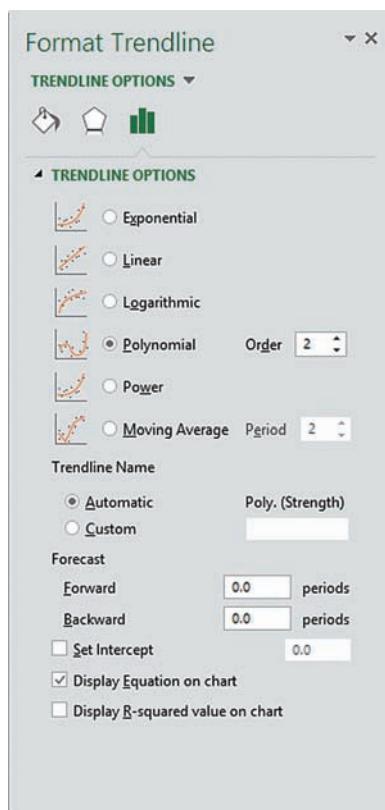
Key Technique Use the exponential operator (^) in a column of formulas to create a quadratic term.

Example Create the quadratic term for the Section 15.1 fly ash percentage analysis and construct the Figure 15.3 scatter plot that shows the quadratic relationship between fly ash percentage and strength shown on page 599.

To create the quadratic term, open to the **DATA worksheet** of the **FlyAsh workbook**. That worksheet contains the independent variable **FlyAsh%** in column A and the dependent variable **Strength** in column B. Select column B (**Strength**), right-click, and click **Insert** from the shortcut menu to add a new column B. (Strength becomes column C.) Enter the label **FlyAsh%^2** in cell B1 and then enter the formula **=A2^2** in cell B2. Copy this formula down the column through all the data rows.

Perform a regression analysis using this new variable by adapting either the Section EG14.1 *PHStat* or *In-Depth Excel* instructions. Then adapt the appropriate Section EG2.5 instructions to construct a scatter plot. Select that chart. Then select **Design** → **Add Chart Element** → **Trendline** → **More Trendline Options** and in the Format Trendline pane (shown below) click **Polynomial** and check **Display Equation on chart** in the Format Trendline pane.

In Excel versions older than Excel 2013, select **Layout** → **Trendline** → **More Trendline Options** and in the Format Trendline dialog box (similar to the Format Trendline Pane), click **Trendline Options** in the left pane and in the Trendline Options right pane, click **Polynomial**, check **Display Equation on chart**, and click **OK**.



While the quadratic term **FlyAsh%^2** could be created in any column, placing independent variables in contiguous columns is a best practice and mandatory if you use the Analysis ToolPak Regression procedure.

EG15.2 USING TRANSFORMATIONS in REGRESSION MODELS

The Square-Root Transformation

To the worksheet that contains your regression data, add a new column of formulas that computes the square root of one of the independent variables to create a square-root transformation. For example, to create a square root transformation in a blank column D for an independent variable in a column C, enter the formula **=SQRT(C2)** in cell D2 of that worksheet and copy the formula down through all data rows. If the rightmost column in the worksheet contains the dependent variable, first select that column, right-click, and click **Insert** from the shortcut menu and place the transformation in that new column.

The Log Transformation

To the worksheet that contains your regression data, add a new column of formulas that computes the common (base 10) logarithm or natural logarithm (base *e*) of one of the independent variables to create a log transformation. For example, to create a common logarithm transformation in a blank column D for a variable in a column C, enter the formula **=LOG(C2)** in cell D2 of that worksheet and copy the formula down through all data rows. To create a natural logarithm transformation in a blank column D for a variable in column C, enter the formula **=LN(C2)** in cell D2 of that worksheet and copy the formula down through all data rows.

If the dependent variable appears in a column to the immediate right of the independent variable being transformed, first select the dependent variable column, right-click, and click **Insert** from the shortcut menu and then place the transformation of the independent variable in that new column.

EG15.3 COLLINEARITY

PHStat To compute the variance inflationary factor (*VIF*), use the “Interpreting the Regression Coefficients” *PHStat* instructions in Section EG14.1 on page 589, but modify step 6 by checking **Variance Inflationary Factor** (*VIF*) before you click **OK**. The *VIF* will appear in cell B9 of the regression results worksheet, immediately following the Regression Statistics area.

In-Depth Excel To compute the variance inflationary factor, first use the “Interpreting the Regression Coefficients” *In-Depth Excel* instructions in Section EG14.1 on page 589 to create regression results worksheets for every combination of independent variables in which one serves as the dependent variable. Then, in each of the regression results worksheets, enter the label *VIF* in cell **A9** and enter the formula **=1/(1 - B5)** in cell **B9** to compute the *VIF*.

EG15.4 MODEL BUILDING

The Stepwise Regression Approach to Model Building

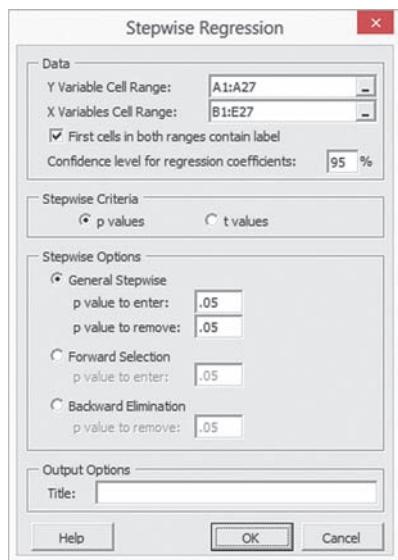
Key Technique Use PHStat to perform a stepwise analysis.

Example Perform the Figure 15.11 stepwise analysis for the standby-hours data that is shown on page 611.

PHStat Use Stepwise Regression.

For the example, open to the **DATA worksheet** of the **Standby workbook**. Select **PHStat → Regression → Stepwise Regression**. In the procedure's dialog box (shown below):

1. Enter **A1:A27** as the **Y Variable Cell Range**.
2. Enter **B1:E27** as the **X Variables Cell Range**.
3. Check **First cells in both ranges contain label**.
4. Enter **95** as the **Confidence level for regression coefficients**.
5. Click **p values** as the **Stepwise Criteria**.
6. Click **General Stepwise** and keep the pair of **.05** values as the **p value to enter** and the **p value to remove**.
7. Enter a **Title** and click **OK**.



This procedure may take more than a few seconds to construct its results. The procedure finishes when the statement “Stepwise ends” is added to the stepwise regression results worksheet (shown in row 29 in Figure 15.11 on page 611).

The Best-Subsets Approach to Model Building

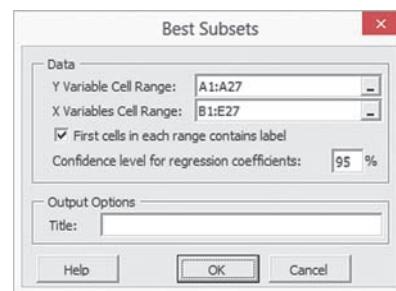
Key Technique Use PHStat to perform a stepwise analysis.

Example Perform the Figure 15.12 best subsets analysis for the standby-hours data that is shown on page 612.

PHStat Use Best Subsets.

For the example, open to the **DATA worksheet** of the **Standby workbook**. Select **PHStat → Regression → Best Subsets**. In the procedure's dialog box (shown below):

1. Enter **A1:A27** as the **Y Variable Cell Range**.
2. Enter **B1:E27** as the **X Variables Cell Range**.
3. Check **First cells in each range contains label**.
4. Enter **95** as the **Confidence level for regression coefficients**.
5. Enter a **Title** and click **OK**.



This procedure constructs many regression results worksheets (that may be seen as a flickering in the Excel window) as it evaluates each subset of independent variables.

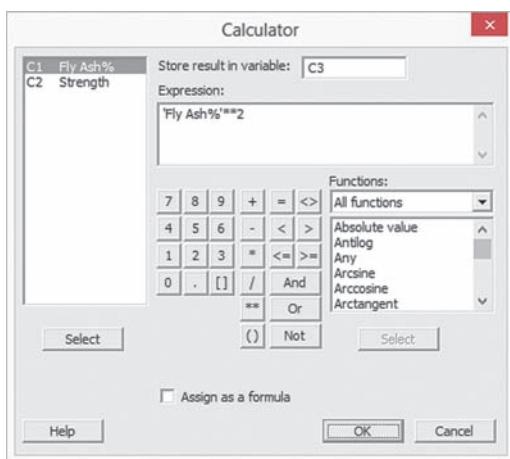
CHAPTER 15 MINITAB GUIDE

MG15.1 The QUADRATIC REGRESSION MODEL

Use **Calculator** to compute the square of one of the independent variables to create a quadratic term.

For example, to create a quadratic term for the Section 15.1 fly ash analysis, open to the **FlyAsh worksheet**. Select **Calc → Calculator**. In the Calculator dialog box (shown on page 627):

1. Enter **C3** in the **Store result in variable** box and press **Tab**.
2. Double-click **C1 Fly Ash%** in the variables list to add 'Fly Ash%' to the **Expression** box.
3. Click ****** and then **2** on the simulated calculator keypad to add ****2** to the **Expression** box.
4. Click **OK**.
5. Enter **Fly Ash%^2** as the name for column **C3**.



To perform a regression analysis using this new variable, use the Section MG14.1 instructions on page 592.

MG15.2 USING TRANSFORMATIONS in REGRESSION MODELS

Use **Calculator** to transform a variable. Open to the worksheet that contains your regression data. Select **Calc → Calculator**. In the Calculator dialog box:

1. Enter the name of the empty column that will contain the transformed values in the **Store result in variable** box and press **Tab**.
2. Select **All functions** from the **Functions** drop-down list.
3. In the list of functions, select one of these choices: **Square root**, **Log base 10**, or **Natural log (log base e)**. Selecting these choices enters **SQRT(number)**, **LOGTEN(number)**, or **LN(number)**, respectively, in the **Expression** box.
4. Double-click the name of the variable to be transformed in the variables list to replace **number** with the variable name in the **Expression** box.
5. Click **OK**.
6. Enter a column name for the transformed values.

To perform a regression analysis using this new variable, see Section MG14.1 on page 592.

MG15.3 COLLINEARITY

To compute the variance inflationary factor, modify the Section MG14.1 “Interpreting the Regression Coefficients” instructions on page 592. In step 15, check **Variance inflation factors** while clearing the other **Display** and **Lack of Fit Test** check boxes in the Regression - Options dialog box.

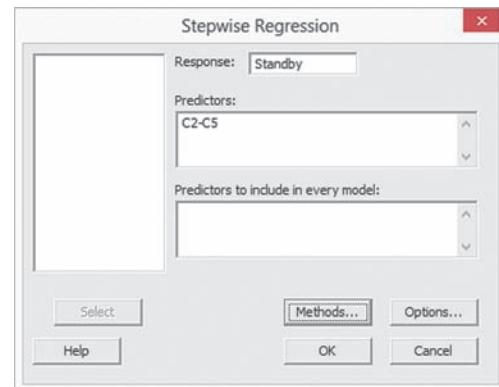
MG15.4 MODEL BUILDING

The Stepwise Regression Approach to Model Building

Use **Stepwise**.

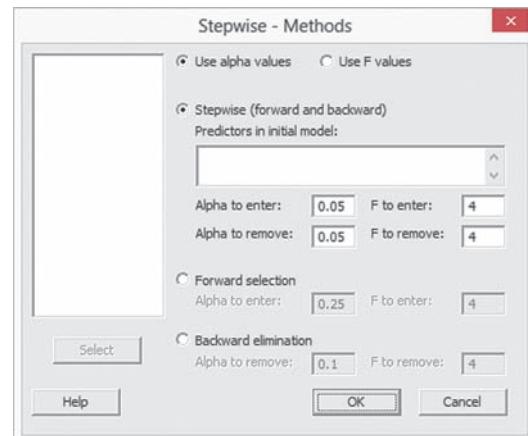
For example, to create the Figure 15.11 stepwise analysis of the standby-hours data on page 611, open to the **Standby** worksheet. Select **Stat → Regression → Stepwise**. In the Stepwise Regression dialog box (shown below):

1. Double-click **C1 Standby** in the variables list to add **Standby** in the **Response** box.
2. Enter **C2-C5** in the **Predictors** box. (Entering **C2-C5** is a shortcut way of referring to the four variables in columns 2 through 5. This shortcut avoids having to double-click the name of each of these variables in order to add them to the Predictors box.)
3. Click **Methods**.



In the Stepwise-Methods dialog box (shown below):

4. Click **Use alpha values**.
5. Click **Stepwise**.
6. Enter **0.05** in the **Alpha to enter** box and **0.05** in the **Alpha to remove** box.
7. Click **OK**.



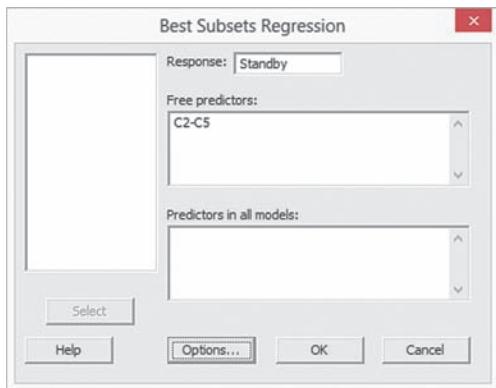
8. Back in the Stepwise Regression dialog box, click **OK**.

The Best-Subsets Approach to Model Building

Use Best Subsets.

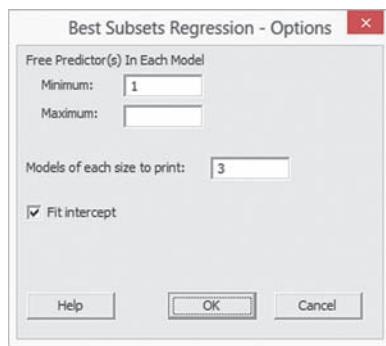
For example, to create the Figure 15.12 stepwise analysis of the standby-hours data on page 612, open to the **Standby worksheet**. Select **Stat → Regression → Best Subsets**. In the Best Subsets Regression dialog box (shown below):

1. Double-click **C1 Standby** in the variables list to add **Standby** in the **Response** box.
2. Enter **C2-C5** in the **Free Predictors** box. (Entering **C2-C5** is a shortcut way of referring to the four variables in columns 2 through 5 as explained in the previous set of instructions.)
3. Click **Options**.



In the Best Subsets Regression - Options dialog box (shown below):

4. Enter **1** in the **Minimum** box and keep the **Maximum** box empty.
5. Enter **3** in the **Models of each size to print** box.
6. Check **Fit intercept**
7. Click **OK**.
8. Back in the Best Subsets Regression dialog box, click **OK**.



CHAPTER 16

Time-Series Forecasting

CONTENTS

- 16.1 The Importance of Business Forecasting
- 16.2 Component Factors of Time-Series Models
- 16.3 Smoothing an Annual Time Series
- 16.4 Least-Squares Trend Fitting and Forecasting
- 16.5 Autoregressive Modeling for Trend Fitting and Forecasting
- 16.6 Choosing an Appropriate Forecasting Model
- 16.7 Time-Series Forecasting of Seasonal Data
- 16.8 Index Numbers (*online*)

THINK ABOUT THIS: Let The Model User Beware

USING STATISTICS: Principled Forecasting, Revisited

CHAPTER 16 EXCEL GUIDE

CHAPTER 16 MINITAB GUIDE

OBJECTIVES

To learn to construct different time-series forecasting models—moving averages, exponential smoothing, the linear trend, the quadratic trend, the exponential trend—and the autoregressive models and least-squares models for seasonal data

To Learn to choose the most appropriate time-series forecasting model

USING STATISTICS

Principled Forecasting

You are a financial analyst for The Principled, a large financial services company. You need to better evaluate investment opportunities for your clients. To assist in the forecasting, you have collected a time series for yearly movie attendance and the revenues of two large well-known companies, The Coca-Cola Company, and Wal-Mart Stores, Inc. Each time series has unique characteristics due to differences in business activities and growth patterns. You understand that you can choose among several different types of forecasting models. How do you decide which type of forecasting is best? How do you use the information gained from the forecasting models to evaluate investment opportunities for your clients?



In Chapters 13 through 15, you used regression analysis as a tool for model building and prediction. In this chapter, regression analysis and other statistical methodologies are applied to time-series data. A **time series** is a set of numerical data collected over time. Due to differences in the features of data for various investments described in the Using Statistics scenario, you need to consider several different approaches for forecasting time-series data.

This chapter begins with an introduction to the importance of business forecasting (see Section 16.1) and a description of the components of time-series models (see Section 16.2). The coverage of forecasting models begins with annual time-series data. Section 16.3 presents moving averages and exponential smoothing methods for smoothing a series. This is followed by least-squares trend fitting and forecasting in Section 16.4 and autoregressive modeling in Section 16.5. Section 16.6 discusses how to choose among alternative forecasting models. Section 16.7 develops models for monthly and quarterly time series.

16.1 The Importance of Business Forecasting

Because economic and business conditions vary over time, managers must find ways to keep abreast of the effects that such changes will have on their organizations. One technique that can aid in planning for the future needs is *forecasting*. **Forecasting** involves monitoring changes that occur over time and predicting the future. For example, marketing executives at a retailer might forecast product demand, sales revenues, consumer preferences, and inventory, among other things, in order to make decisions regarding product promotions and strategic planning. Government officials forecast unemployment, inflation, industrial production, and revenues from income taxes in order to formulate economic policies. And the administrators of a college need to forecast student enrollment in order to plan for the construction of dormitories and academic facilities and plan for student and faculty recruitment.

Two common approaches to forecasting are *qualitative* and *quantitative* forecasting. **Qualitative forecasting methods** are especially important when historical data are unavailable. Qualitative forecasting methods are considered to be highly subjective and judgmental. **Quantitative forecasting methods** make use of historical data. The goal of these methods is to use past data to predict future values.

Quantitative forecasting methods are subdivided into two types: *time series* and *causal*. **Time-series forecasting methods** involve forecasting future values based entirely on the past and present values of a variable. For example, the daily closing prices of a particular stock on the New York Stock Exchange constitute a time series. Other examples of economic or business time series are the consumer price index (CPI), the quarterly gross domestic product (GDP), and the annual sales revenues of a particular company.

Causal forecasting methods involve the determination of factors that relate to the variable you are trying to forecast. These include multiple regression analysis with lagged variables, econometric modeling, leading indicator analysis, and other economic barometers that are beyond the scope of this text (see references 2–4). The emphasis in this chapter is on time-series forecasting methods.

16.2 Component Factors of Time-Series Models

Time-series forecasting assumes that the factors that have influenced activities in the past and present will continue to do so in approximately the same way in the future. Time-series forecasting seeks to identify and isolate these component factors in order to make predictions. Typically, the following four factors are examined in time-series models:

- Trend
- Cyclical effect
- Irregular or random effect
- Seasonal effect

A **trend** is an overall long-term upward or downward movement in a time series. The **cyclical effect** involves the up-and-down swings or movements through the series. Cyclical movements vary in length, usually lasting from 2 to 10 years. They differ in intensity and are often correlated with a business cycle. In some time periods, the values are higher than would be predicted by a trend line (i.e., they are at or near the peak of a cycle). In other time periods, the values are lower than would be predicted by a trend line (i.e., they are at or near the bottom of a cycle). Any data that do not follow the trend modified by the cyclical component are considered part of the **irregular effect**, or **random effect**. When you have monthly or quarterly data, an additional component, the **seasonal effect**, is considered, along with the trend, cyclical, and irregular effects.

Your first step in a time-series analysis is to visualize the data and observe whether any patterns exist over time. You must determine whether there is a long-term upward or downward movement in the series (i.e., a trend). If there is no obvious long-term upward or downward trend, then you can use moving averages or exponential smoothing to smooth the series (see Section 16.3). If a trend is present, you can consider several time-series forecasting methods. (See Sections 16.4 and 16.5 for forecasting annual data and Section 16.7 for forecasting monthly or quarterly time series.)

16.3 Smoothing an Annual Time Series

One of the investments considered in The Principled scenario is the entertainment industry. Table 16.1 presents the yearly U.S. and Canada movie attendance (in billions) from 2002 through 2012 (stored in **Movie Attendance**) and Figure 16.1 presents the time-series plot of these data.

TABLE 16.1

Movie Attendance from 2002 Through 2012

Year	Attendance (billions)	Year	Attendance (billions)	Year	Attendance (billions)
2002	1.57	2006	1.40	2010	1.34
2003	1.52	2007	1.40	2011	1.28
2004	1.50	2008	1.34	2012	1.36
2005	1.38	2009	1.42		

Source: Data extracted from *Theatrical Market Statistics*, 2011 and 2012 editions, Motion Picture Association of America, Inc.

FIGURE 16.1

Time-series plot of movie attendance from 2002 through 2012

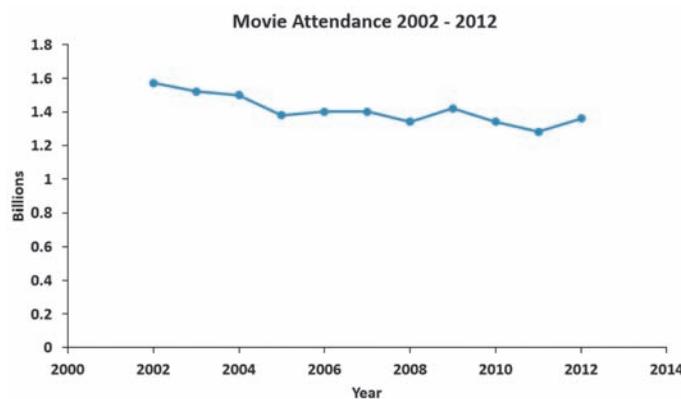


Figure 16.1 seems to show a slight downward trend in movie attendance between 2002 to 2012. However, when you examine annual data such as in Figure 16.1, your visual impression of the long-term trend in the series is sometimes obscured by the amount of variation from year to year. Often, you cannot judge whether any long-term upward or downward trend exists in the series. To get a better overall impression of the pattern of movement in the data over time, you can use the methods of *moving averages* or *exponential smoothing*.

Moving Averages

Moving averages for a chosen period of length L consist of a series of means, each computed over time for a sequence of L observed values. Moving averages, represented by the symbol $MA(L)$, can be greatly affected by the value chosen for L , which should be an integer value that corresponds to, or is a multiple of, the estimated average length of a cycle in the time series.

To illustrate, suppose you want to compute five-year moving averages from a series that has $n = 11$ years. Because $L = 5$, the five-year moving averages consist of a series of means computed by averaging consecutive sequences of five values. You compute the first five-year moving average by summing the values for the first five years in the series and dividing by 5:

$$MA(5) = \frac{Y_1 + Y_2 + Y_3 + Y_4 + Y_5}{5}$$

You compute the second five-year moving average by summing the values of years 2 through 6 in the series and then dividing by 5:

$$MA(5) = \frac{Y_2 + Y_3 + Y_4 + Y_5 + Y_6}{5}$$

You continue this process until you have computed the last of these five-year moving averages by summing the values of the last 5 years in the series (i.e., years 7 through 11) and then dividing by 5:

$$MA(5) = \frac{Y_7 + Y_8 + Y_9 + Y_{10} + Y_{11}}{5}$$

When you have annual time-series data, L should be an *odd* number of years. By following this rule, you are unable to compute any moving averages for the first $(L - 1)/2$ years or the last $(L - 1)/2$ years of the series. Thus, for a five-year moving average, you cannot make computations for the first two years or the last two years of the series.

When plotting moving averages, you plot each of the computed values against the middle year of the sequence of years used to compute it. If $n = 11$ and $L = 5$, the first moving average is centered on the third year, the second moving average is centered on the fourth year, and the last moving average is centered on the ninth year. Example 16.1 illustrates the computation of five-year moving averages.

Student Tip

Remember that you cannot compute moving averages at the beginning and at the end of the series.

EXAMPLE 16.1

Computing Five-Year Moving Averages

The following data represent total revenues (in \$millions) for a fast-food store over the 11-year period 2003 to 2013:

4.0 5.0 7.0 6.0 8.0 9.0 5.0 7.0 7.5 5.5 6.5

Compute the five-year moving averages for this annual time series and plot the revenue and moving averages.

SOLUTION To compute the five-year moving averages, you first compute the total for the five years and then divide this total by 5. The first of the five-year moving averages is

$$MA(5) = \frac{Y_1 + Y_2 + Y_3 + Y_4 + Y_5}{5} = \frac{4.0 + 5.0 + 7.0 + 6.0 + 8.0}{5} = \frac{30.0}{5} = 6.0$$

The moving average is centered on the middle value—the third year of this time series. To compute the second of the five-year moving averages, you compute the total of the second through sixth years and divide this total by 5:

$$MA(5) = \frac{Y_2 + Y_3 + Y_4 + Y_5 + Y_6}{5} = \frac{5.0 + 7.0 + 6.0 + 8.0 + 9.0}{5} = \frac{35.0}{5} = 7.0$$

This moving average is centered on the new middle value—the fourth year of the time series. The remaining moving averages are

$$MA(5) = \frac{Y_3 + Y_4 + Y_5 + Y_6 + Y_7}{5} = \frac{7.0 + 6.0 + 8.0 + 9.0 + 5.0}{5} = \frac{35.0}{5} = 7.0$$

$$MA(5) = \frac{Y_4 + Y_5 + Y_6 + Y_7 + Y_8}{5} = \frac{6.0 + 8.0 + 9.0 + 5.0 + 7.0}{5} = \frac{35.0}{5} = 7.0$$

$$MA(5) = \frac{Y_5 + Y_6 + Y_7 + Y_8 + Y_9}{5} = \frac{8.0 + 9.0 + 5.0 + 7.0 + 7.5}{5} = \frac{36.5}{5} = 7.3$$

$$MA(5) = \frac{Y_6 + Y_7 + Y_8 + Y_9 + Y_{10}}{5} = \frac{9.0 + 5.0 + 7.0 + 7.5 + 5.5}{5} = \frac{34.0}{5} = 6.8$$

$$MA(5) = \frac{Y_7 + Y_8 + Y_9 + Y_{10} + Y_{11}}{5} = \frac{5.0 + 7.0 + 7.5 + 5.5 + 6.5}{5} = \frac{31.5}{5} = 6.3$$

These moving averages are centered on their respective middle values—the fifth, sixth, seventh, eighth, and ninth years in the time series. When you use the five-year moving averages, you are unable to compute a moving average for the first two or last two values in the time series.

Figure 16.2 plots the revenues and the five-year moving average for the fast-food store. Observe that the moving average exhibits much less variation than the revenues since they have smoothed the data.

FIGURE 16.2

Fast-food store revenue and five-year moving average

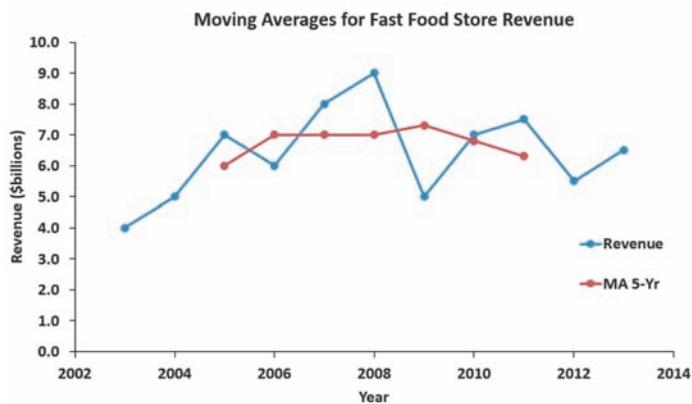
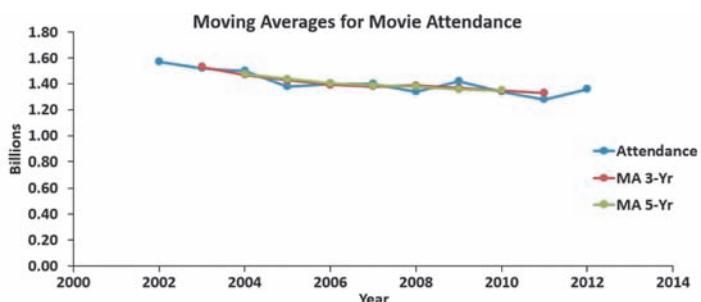


FIGURE 16.3

Excel three- and five-year moving averages and plot for the movie attendance data

A	B	C	D	
1	Year	Attendance	MA 3-Yr	MA 5-Yr
2	2002	1.57	#N/A	#N/A
3	2003	1.52	1.5300	#N/A
4	2004	1.50	1.4667	1.4740
5	2005	1.38	1.4267	1.4400
6	2006	1.40	1.3933	1.4040
7	2007	1.40	1.3800	1.3880
8	2008	1.34	1.3867	1.3800
9	2009	1.42	1.3667	1.3560
10	2010	1.34	1.3467	1.3480
11	2011	1.28	1.3267	#N/A
12	2012	1.36	#N/A	#N/A



In Figure 16.3, there is no three-year moving average for the first year and the last year, and there is no five-year moving average for the first two years and last two years. Both the three-year and five-year moving averages have smoothed out the variation that exists in the movie attendance. Both the three-year and five-year moving average show a slight downward trend in movie attendance. The five-year moving average smooths the series more than the three-year moving average because the period is longer. However, the longer the period, the smaller the number of moving averages you can compute. Therefore, selecting moving averages that are longer than five or seven years is usually undesirable because too many moving average values are missing at the beginning and end of the series.

The selection of L , the length of the period used for constructing the averages, is highly subjective. If cyclical fluctuations are present in the data, you should choose an integer value of L that corresponds to (or is a multiple of) the estimated length of a cycle in the series. For annual time-series data that has no obvious cyclical fluctuations, you should choose three years, five years, or seven years as the value of L , depending on the amount of smoothing desired and the amount of data available.

Exponential Smoothing

Exponential smoothing consists of a series of *exponentially weighted* moving averages. The weights assigned to the values change so that the most recent (the last) value receives the highest weight, the previous value receives the second-highest weight, and so on, with the first value receiving the lowest weight. Throughout the series, each exponentially smoothed value depends on all previous values, which is an advantage of exponential smoothing over the method of moving averages. Exponential smoothing also allows you to compute short-term (one period into the future) forecasts when the presence and type of long-term trend in a time series is difficult to determine.

The equation developed for exponentially smoothing a series in any time period, i , is based on only three terms—the current value in the time series, Y_i ; the previously computed exponentially smoothed value, E_{i-1} ; and an assigned weight or smoothing coefficient, W . You use Equation (16.1) to exponentially smooth a time series.

COMPUTING AN EXPONENTIALLY SMOOTHED VALUE IN TIME PERIOD i

$$\begin{aligned} E_1 &= Y_1 \\ E_i &= WY_i + (1 - W)E_{i-1} \quad i = 2, 3, 4, \dots \end{aligned} \tag{16.1}$$

where

E_i = value of the exponentially smoothed series being computed in time period i

E_{i-1} = value of the exponentially smoothed series already computed in time period $i - 1$

Y_i = observed value of the time series in period i

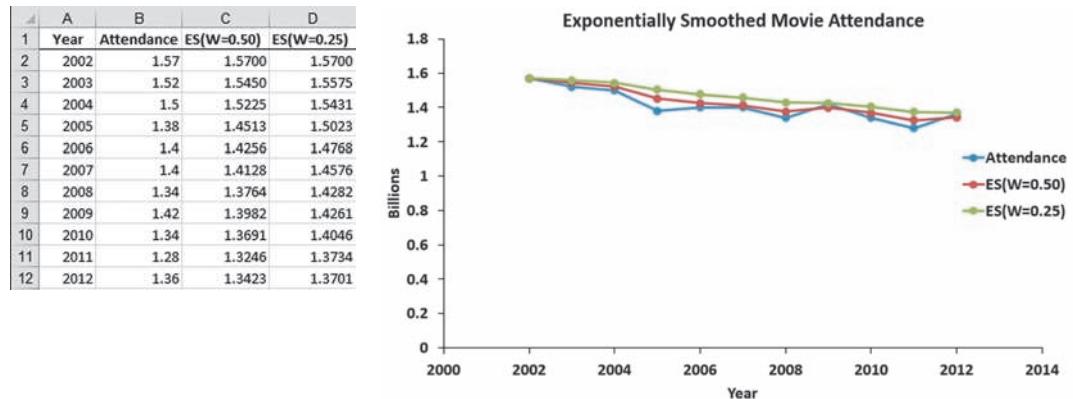
W = subjectively assigned weight or smoothing coefficient
(where $0 < W < 1$); although W can approach 1.0, in virtually all business applications, $W \leq 0.5$

Choosing the weight or smoothing coefficient (i.e., W) that you assign to the time series is critical. Unfortunately, this selection is somewhat subjective. If your goal is to smooth a series by eliminating unwanted cyclical and irregular variations in order to see the overall

long-term tendency of the series, you should select a small value for W (close to 0). If your goal is forecasting future short-term directions, you should choose a large value for W (close to 0.5). Figure 16.4 presents the exponentially smoothed values (with smoothing coefficients $W = 0.50$ and $W = 0.25$), the movie attendance from 2002 to 2012, and a plot of the original data and the two exponentially smoothed time series. Observe that exponential smoothing has smoothed out some of the variation in the movie attendance.

FIGURE 16.4

Exponentially smoothed series ($W = 0.50$ and $W = 0.25$) worksheet and plot for the movie attendance data



To illustrate these exponential smoothing computations for a smoothing coefficient of $W = 0.25$, you begin with the initial value $Y_{2002} = 1.57$ as the first smoothed value ($E_{2002} = 1.57$). Then, using the value of the time series for 2003 ($Y_{2003} = 1.52$), you smooth the series for 2003 by computing

$$\begin{aligned} E_{2003} &= WY_{2003} + (1 - W)E_{2002} \\ &= (0.25)(1.52) + (0.75)(1.57) = 1.5575 \end{aligned}$$

To smooth the series for 2004:

$$\begin{aligned} E_{2004} &= WY_{2004} + (1 - W)E_{2003} \\ &= (0.25)(1.5) + (0.75)(1.5575) = 1.5431 \end{aligned}$$

To smooth the series for 2005:

$$\begin{aligned} E_{2005} &= WY_{2005} + (1 - W)E_{2004} \\ &= (0.25)(1.38) + (0.75)(1.5431) = 1.5023 \end{aligned}$$

You continue this process until you have computed the exponentially smoothed values for all 11 years in the series, as shown in Figure 16.4.

In general, you compute the current smoothed value as follows:

$$\text{Current smoothed value} = (W)(\text{Current value}) + (1 - W)(\text{Previous smoothed value})$$

Remember that the smoothed value for the first year is the observed value in the first year.

To use exponential smoothing for forecasting, you use the smoothed value in the current time period as the forecast of the value in the following period (\hat{Y}_{i+1}).

FORECASTING TIME PERIOD $i + 1$

$$\hat{Y}_{i+1} = E_i \quad (16.2)$$

To forecast the movie attendance in 2013, using a smoothing coefficient of $W = 0.25$, you use the smoothed value for 2012 as its estimate.

$$\begin{aligned}\hat{Y}_{2012+1} &= E_{2012} \\ \hat{Y}_{2013} &= E_{2012} \\ \hat{Y}_{2013} &= 1.3701\end{aligned}$$

The exponentially smoothed forecast for 2013 is 1.3701 billion.

Problems for Section 16.3

LEARNING THE BASICS

16.1 If you are using exponential smoothing for forecasting an annual time series of revenues, what is your forecast for next year if the smoothed value for this year is \$32.4 million?

16.2 Consider a nine-year moving average used to smooth a time series that was first recorded in 1984.

- a. Which year serves as the first centered value in the smoothed series?
- b. How many years of values in the series are lost when computing all the nine-year moving averages?

16.3 You are using exponential smoothing on an annual time series concerning total revenues (in \$millions). You decide to use a smoothing coefficient of $W = 0.20$, and the exponentially smoothed value for 2013 is $E_{2013} = (0.20)(12.1) + (0.80)(9.4)$.

- a. What is the smoothed value of this series in 2013?
- b. What is the smoothed value of this series in 2014 if the value of the series in that year is \$11.5 million?

APPLYING THE CONCEPTS

SELF Test **16.4** The data below (stored in **Drive-ThruSpeed**) represent the average time (in seconds) it took to be served at the drive-through at McDonald's from 1998 to 2012:

Year	Drive-Through Speed (seconds)	Year	Drive-Through Speed (seconds)
1998	177.59	2006	163.90
1999	167.02	2007	167.10
2000	169.88	2008	158.77
2001	170.85	2009	174.22
2002	162.72	2010	179.27
2003	156.92	2011	184.20
2004	152.52	2012	188.83
2005	167.90		

Source: Data extracted from bit.ly/QEIeOW.

- a. Plot the time series.
- b. Fit a three-year moving average to the data and plot the results.
- c. Using a smoothing coefficient of $W = 0.50$, exponentially smooth the series and plot the results.
- d. What is your exponentially smoothed forecast for 2013?
- e. Repeat (c) and (d), using $W = 0.25$.
- f. Compare the results of (d) and (e).
- g. What conclusions can you reach about the drive-through speed at McDonald's?

16.5 The following data, stored in **Spills** provide the number of oil spills in the Gulf of Mexico from 1996 to 2012:

Year	Number of Spills	Year	Number of Spills
1996	4	2005	49
1997	3	2006	14
1998	9	2007	4
1999	5	2008	33
2000	7	2009	11
2001	9	2010	5
2002	12	2011	3
2003	12	2012	8
2004	22		

Source: Data extracted from www.bsee.gov/Inspection-and-Enforcement/Accidents-and-Incidents/Spills.

- a. Plot the time series.
- b. Fit a three-year moving average to the data and plot the results.
- c. Using a smoothing coefficient of $W = 0.50$, exponentially smooth the series and plot the results.
- d. What is your exponentially smoothed forecast for 2013?
- e. Repeat (c) and (d), using $W = 0.25$.
- f. Compare the results of (d) and (e).
- g. What conclusions can you reach concerning the number of oil spills in the Gulf of Mexico?

16.6 How have stocks performed in the past? The following table presents the data stored in **Stock Performance**, which show the performance of a broad measure of stock performance (by percentage) for each decade from the 1830s through the 2000s:

Decade	Performance (%)	Decade	Performance (%)
1830s	2.8	1920s	13.3
1840s	12.8	1930s	-2.2
1850s	6.6	1940s	9.6
1860s	12.5	1950s	18.2
1870s	7.5	1960s	8.3
1880s	6.0	1970s	6.6
1890s	5.5	1980s	16.6
1900s	10.9	1990s	17.6
1910s	2.2	2000s*	-0.5

*Through December 15, 2009.

Source: T. Lauricella, "Investors Hope the '10s Beat the '00s," *The Wall Street Journal*, December 21, 2009, pp. C1, C2.

- Plot the time series.
- Fit a three-period moving average to the data and plot the results.
- Using a smoothing coefficient of $W = 0.50$, exponentially smooth the series and plot the results.
- What is your exponentially smoothed forecast for the 2010s?
- Repeat (c) and (d), using $W = 0.25$.
- Compare the results of (d) and (e).
- What conclusions can you reach concerning how stocks have performed in the past?

16.7 The following data (stored in **CoffeePricesPortugal**) represent the retail price of coffee (in €/kg) in Portugal from 2004 to 2011:

Year	Retail Price (€/kg)
2004	8.60
2005	8.53
2006	8.32
2007	8.23
2008	8.58
2009	8.45
2010	8.31
2011	8.60

Source: International Coffee Organization, www.ico.org.

- Plot the data.
- Fit a three-year moving average to the data and plot the results.
- Using a smoothing coefficient of $W = 0.50$, exponentially smooth the series and plot the results.
- What is your exponentially smoothed forecast for 2012?
- Repeat (c) and (d), using a smoothing coefficient of $W = 0.25$.
- Compare the results of (d) and (e).
- What conclusions can you reach about the retail price of coffee in Portugal?

16.8 The file **Audits** contains the number of audits of corporations with assets of more than \$250 million conducted by the Internal Revenue Service. (Data extracted from www.irs.gov.)

- Plot the data.
- Fit a three-year moving average to the data and plot the results.
- Using a smoothing coefficient of $W = 0.50$, exponentially smooth the series and plot the results.
- What is your exponentially smoothed forecast for 2013?
- Repeat (c) and (d), using a smoothing coefficient of $W = 0.25$.
- Compare the results of (d) and (e).
- What conclusions can you reach concerning the number of audits of corporations with assets of more than \$250 million conducted by the Internal Revenue Service?

16.4 Least-Squares Trend Fitting and Forecasting

Trend is the component factor of a time series most often used to make intermediate and long-range forecasts. To get a visual impression of the overall long-term movements in a time series, you construct a time-series plot. If a straight-line trend adequately fits the data, you can use a linear trend model [see Equation (16.3) and Section 13.2]. If the time-series data indicate some long-run downward or upward quadratic movement, you can use a quadratic trend model [see Equation (16.4) and Section 15.1]. When the time-series data increase at a rate such that the percentage difference from value to value is constant, you can use an exponential trend model [see Equation (16.5)].

The Linear Trend Model

The **linear trend model**:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

is the simplest forecasting model. Equation (16.3) defines the linear trend forecasting equation.

LINEAR TREND FORECASTING EQUATION

$$\hat{Y}_i = b_0 + b_1 X_i \quad (16.3)$$

Recall that in linear regression analysis, you use the method of least squares to compute the sample slope, b_1 , and the sample Y intercept, b_0 . You then substitute the values for X into Equation (16.3) to predict Y .

When using the least-squares method for fitting trends in a time series, you can simplify the interpretation of the coefficients by assigning coded values to the X (time) variable. You assign consecutively numbered integers, starting with 0, as the coded values for the time periods. For example, in time-series data that have been recorded annually for 17 years, you assign the coded value 0 to the first year, the coded value 1 to the second year, the coded value 2 to the third year, and so on, concluding by assigning 16 to the seventeenth year.

In The Principled scenario on page 629, one of the companies of interest is The Coca-Cola Company. Founded in 1886 and headquartered in Atlanta, Georgia, Coca-Cola manufactures, distributes, and markets more than 500 beverage brands in over 200 countries worldwide. Brands include Coca-Cola, Diet Coke, Fanta, and Sprite, four of the world's top five nonalcoholic sparkling beverage products. According to The Coca-Cola Company's website, revenues in 2012 topped \$48 billion. Table 16.2 lists The Coca-Cola Company's gross revenues (in \$billions) from 1996 to 2012 (stored in [Coca-Cola](#)).

TABLE 16.2

Revenues for The Coca-Cola Company (1996–2012)

Year	Revenues (\$billions)	Year	Revenues (\$billions)
1996	18.5	2005	23.1
1997	18.9	2006	24.1
1998	18.8	2007	28.9
1999	19.8	2008	31.9
2000	20.5	2009	31.0
2001	20.1	2010	35.1
2002	19.6	2011	46.5
2003	21.0	2012	48.0
2004	21.9		

Source: Data extracted from *Mergent's Handbook of Common Stocks*, 2006; and www.thecocacolacompany.com/investors/annual_other_reports.html.

Figure 16.5 presents the regression results for the simple linear regression model that uses the consecutive coded values 0 through 16 as the X (coded year) variable. These results produce the following linear trend forecasting equation:

$$\hat{Y}_i = 13.2745 + 1.6326X_i$$

where $X_1 = 0$ represents 1996.

FIGURE 16.5

Excel and Minitab regression results for the linear trend model to forecast revenues (in \$billions) for The Coca-Cola Company

Linear Trend Model for The Coca-Cola Company Revenues							Regression Analysis: Revenues versus Coded Year					
Regression Statistics												
1	2	3	4	5	6	7	Predictor	Coeff	SE Coef	T	P	
Multiple R	0.8782	R Square	0.7712	Adjusted R Square	0.7559	Standard Error	4.6381	Constant	13.275	2.154	6.16	0.000
Observations	17						Coded Year	1.6326	0.2296	7.11	0.000	
								S = 4.63814	R-Sq = 77.1%	R-Sq(adj) = 75.6%		
ANOVA							Analysis of Variance					
11	df	SS	MS	F	Significance F		Source	DF	SS	MS	F	P
Regression	1	1087.4736	1087.4736	50.5511	0.0000		Regression	1	1087.5	1087.5	50.55	0.000
Residual	15	322.6853	21.5124				Residual Error	15	322.7		21.5	
Total	16	1410.1588					Total	16	1410.2			
Coefficients												
Intercept	13.2745	2.1540	6.1626	0.0000	8.6833	17.8658						
Coded Year	1.6326	0.2296	7.1099	0.0000	1.1432	2.1220						

You interpret the regression coefficients as follows:

- The Y intercept, $b_0 = 13.2745$, is the predicted revenues (in \$billions) at The Coca-Cola Company during the origin, or base, year, 1996.
- The slope, $b_1 = 1.6326$, indicates that revenues are predicted to increase by \$1.6326 billion per year.

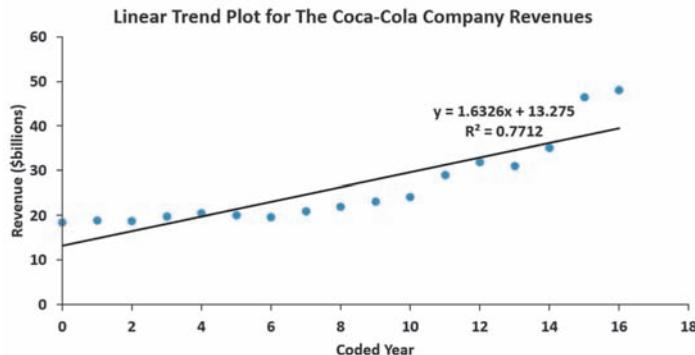
To project the trend in the revenues at Coca-Cola to 2013, you substitute $X_{18} = 17$, the code for 2013 into the linear trend forecasting equation:

$$\hat{Y}_i = 13.2745 + 1.6326(17) = 41.0287 \text{ billions of dollars}$$

The trend line is plotted in Figure 16.6, along with the observed values of the time series. There is a strong upward linear trend, and r^2 is 0.7712, indicating that more than 77% of the variation in revenues is explained by the linear trend of the time series. However, you can observe that the revenue for the most recent year, 2012, is substantially above the trend line, that the early years are also slightly above the trend line, but the middle years are below the trend line. To investigate whether a different trend model might provide a better fit, a *quadratic* trend model and an *exponential* trend model can be fitted.

FIGURE 16.6

Plot of the linear trend forecasting equation for The Coca-Cola Company revenue data



The Quadratic Trend Model

A **quadratic trend model**:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

is a nonlinear model that contains a linear term and a curvilinear term in addition to a Y intercept. Using the least-squares method for a quadratic model described in Section 15.1, you can develop a quadratic trend forecasting equation, as presented in Equation (16.4).

QUADRATIC TREND FORECASTING EQUATION

$$\hat{Y}_i = b_0 + b_1 X_i + b_2 X_i^2 \quad (16.4)$$

where

b_0 = estimated Y intercept

b_1 = estimated *linear* effect on Y

b_2 = estimated *quadratic* effect on Y

Figure 16.7 presents the regression results for the quadratic trend model used to forecast revenues at The Coca-Cola Company.

FIGURE 16.7

Excel and Minitab regression results for the quadratic trend model to forecast revenues (in \$billions) for The Coca-Cola Company

A	B	C	D	E	F	G
1 Quadratic Trend Model for The Coca-Cola Company Revenues						
2						
3 Regression Statistics						
4 Multiple R	0.9782					
5 R Square	0.9568					
6 Adjusted R Square	0.9506					
7 Standard Error	2.0860					
8 Observations	17					
9						
10 ANOVA						
11	df	SS	MS	F	Significance F	
12 Regression	2	1349.2383	674.6191	155.0326	0.0000	
13 Residual	14	60.9205	4.3515			
14 Total	16	1410.1588				
15						
16	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17 Intercept	20.6249	1.3552	15.2186	0.0000	17.7182	23.5316
18 Coded Year	-1.3075	0.3929	-3.3280	0.0050	-2.1502	-0.4649
19 Coded Year Squared	0.1838	0.0237	7.7560	0.0000	0.1329	0.2346

Regression Analysis: Revenues versus Coded Year, Coded Year Squared						
The regression equation is						
Revenues = 20.6 - 1.31 Coded Year + 0.184 Coded Year Squared						
Predictor	Coeff	SB Coef	T	P		
Constant	20.625	1.355	15.22	0.000		
Coded Year	-1.3075	0.3929	-3.33	0.005		
Coded Year Squared	0.18376	0.02369	7.76	0.000		
S = 2.08602	R-Sq = 95.7%	R-Sq(adj) = 95.1%				
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	2	1349.24	674.62	155.03	0.000	
Residual Error	14	60.92	4.35			
Total	16	1410.16				

In Figure 16.7,

$$\hat{Y}_i = 20.6249 - 1.3075X_i + 0.1838X_i^2$$

where the year coded 0 is 1996.

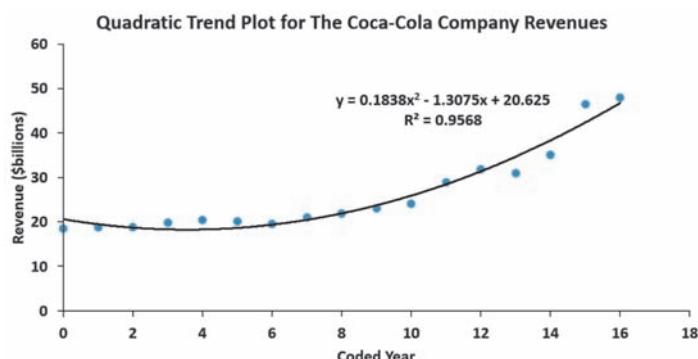
To compute a forecast using the quadratic trend equation, you substitute the appropriate coded X value into this equation. For example, to forecast the trend in revenues for 2013 (i.e., $X = 17$),

$$\hat{Y}_i = 20.6249 - 1.3075(17) + 0.1838(17)^2 = 51.5156$$

Figure 16.8 plots the quadratic trend forecasting equation along with the time series for the actual data. This quadratic trend model provides a better fit ($\text{adjusted } r^2 = 0.9506$) to the time series than does the linear trend model. The t_{STAT} test statistic for the contribution of the quadratic term to the model is 7.756 ($p\text{-value} = 0.0000$).

FIGURE 16.8

Plot of the quadratic trend forecasting equation for The Coca-Cola Company revenue data



The Exponential Trend Model

When a time series increases at a rate such that the percentage difference from value to value is constant, an exponential trend is present. Equation (16.5) defines the **exponential trend model**.

EXPONENTIAL TREND MODEL

$$Y_i = \beta_0 \beta_1^{X_i} \epsilon_i \quad (16.5)$$

where

$$\begin{aligned}\beta_0 &= Y \text{ intercept} \\ (\beta_1 - 1) \times 100\% &= \text{annual compound growth rate (in \%)}\end{aligned}$$

¹Alternatively, you can use base e logarithms. For more information on logarithms, see Section A.3 in Appendix A.

The model in Equation (16.5) is not in the form of a linear regression model. To transform this nonlinear model to a linear model, you use a base 10 logarithm transformation.¹ Taking the logarithm of each side of Equation (16.5) results in Equation (16.6).

TRANSFORMED EXPONENTIAL TREND MODEL

$$\begin{aligned}\log(Y_i) &= \log(\beta_0 \beta_1^{X_i} \epsilon_i) \\ &= \log(\beta_0) + \log(\beta_1^{X_i}) + \log(\epsilon_i) \\ &= \log(\beta_0) + X_i \log(\beta_1) + \log(\epsilon_i)\end{aligned} \quad (16.6)$$

Student Tip

Log is the symbol used for base 10 logarithms. The log of a number is the power that 10 needs to be raised to equal that number.

Equation (16.6) is a linear model you can estimate using the least-squares method, with $\log(Y_i)$ as the dependent variable and X_i as the independent variable. This results in Equation (16.7).

EXPONENTIAL TREND FORECASTING EQUATION

$$\log(\hat{Y}_i) = b_0 + b_1 X_i \quad (16.7a)$$

where

$$\begin{aligned}b_0 &= \text{estimate of } \log(\beta_0) \text{ and thus } 10^{b_0} = \hat{\beta}_0 \\ b_1 &= \text{estimate of } \log(\beta_1) \text{ and thus } 10^{b_1} = \hat{\beta}_1\end{aligned}$$

therefore,

$$\hat{Y}_i = \hat{\beta}_0 \hat{\beta}_1^{X_i} \quad (16.7b)$$

where

$$(\hat{\beta}_1 - 1) \times 100\% \text{ is the estimated annual compound growth rate (in \%)}$$

Figure 16.9 shows the Excel and Minitab regression results for an exponential trend model of revenues at The Coca-Cola Company.

Using Equation (16.7a) and the results from Figure 16.9,

$$\log(\hat{Y}_i) = 1.2005 + 0.0248 X_i$$

where the year coded 0 is 1996.

FIGURE 16.9

Excel and Minitab regression results for the exponential trend model to forecast revenues (in \$billions) for The Coca-Cola Company

A	B	C	D	E	F	G	
1 Exponential Trend Model for The Coca-Cola Company Revenues							
2							
3 Regression Statistics							
4	Multiple R	0.9233					
5	R Square	0.8525					
6	Adjusted R Square	0.8427					
7	Standard Error	0.0538					
8	Observations	17					
9							
10 ANOVA							
11	df	SS	MS	F	Significance F		
12	Regression	1	0.2506	0.2506	86.7077	0.0000	
13	Residual	15	0.0434	0.0029			
14	Total	16	0.2940				
15							
16	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	
17	Intercept	1.2005	0.0250	48.0827	0.0000	1.1473	1.2538
18	Coded Year	0.0248	0.0027	9.3117	0.0000	0.0191	0.0305

Predictor	Coef	SE Coef	T	P
Constant	1.20055	0.02497	48.08	0.000
Coded Year	0.024784	0.002662	9.31	0.000
S	0.0537625	R-Sq = 85.3%	R-Sq(adj) = 84.3%	

Source	DF	SS	MS	F	P
Regression	1	0.25062	0.25062	86.71	0.000
Residual Error	15	0.04336	0.00289		
Total	16	0.29398			

You compute the values for $\hat{\beta}_0$ and $\hat{\beta}_1$ by taking the antilog of the regression coefficients (b_0 and b_1):

$$\begin{aligned}\hat{\beta}_0 &= \text{antilog}(b_0) = \text{antilog}(1.2005) = 10^{1.2005} = 15.8672 \\ \hat{\beta}_1 &= \text{antilog}(b_1) = \text{antilog}(0.0248) = 10^{0.0248} = 1.0588\end{aligned}$$

Thus, using Equation (16.7b), the exponential trend forecasting equation is

$$\hat{Y}_i = (15.8672)(1.0588)^{X_i}$$

where the year coded 0 is 1996.

The Y intercept, $\hat{\beta}_0 = 15.8672$ billions of dollars, is the revenue forecast for the base year 1996. The value $(\hat{\beta}_1 - 1) \times 100\% = 5.88\%$ is the annual compound growth rate in revenues at The Coca-Cola Company.

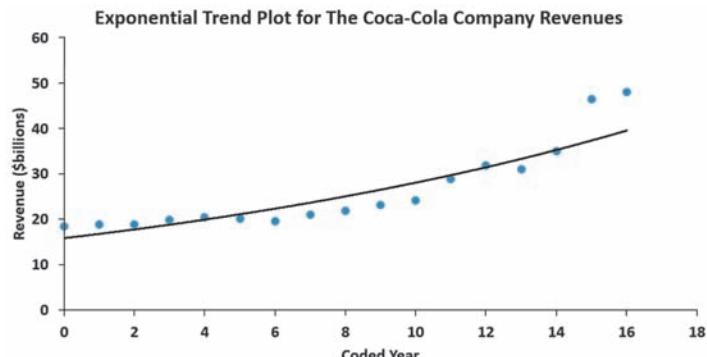
For forecasting purposes, you substitute the appropriate coded X values into either Equation (16.7a) or Equation (16.7b). For example, to forecast revenues for 2013 (i.e., $X = 17$) using Equation (16.7a),

$$\begin{aligned}\log(\hat{Y}_i) &= 1.2005 + 0.0248(17) = 1.6221 \\ \hat{Y}_i &= \text{antilog}(1.6221) = 10^{1.6221} = 41.889 \text{ billions of dollars}\end{aligned}$$

Figure 16.10 plots the exponential trend forecasting equation, along with the time-series data. The adjusted r^2 for the exponential trend model (0.8427) is greater than the adjusted r^2 for the linear trend model (0.7559) but less than for the quadratic model (0.9506).

FIGURE 16.10

Plot of the exponential trend forecasting equation for The Coca-Cola Company revenue data



Model Selection Using First, Second, and Percentage Differences

You have used the linear, quadratic, and exponential models to forecast revenues for The Coca-Cola Company. How can you determine which of these models is the most appropriate model? In addition to visually inspecting time-series plots and comparing adjusted r^2 values, you can compute and examine first, second, and percentage differences. The identifying features of linear, quadratic, and exponential trend models are as follows:

- If a linear trend model provides a perfect fit to a time series, then the first differences are constant. Thus,

$$(Y_2 - Y_1) = (Y_3 - Y_2) = \dots = (Y_n - Y_{n-1})$$

- If a quadratic trend model provides a perfect fit to a time series, then the second differences are constant. Thus,

$$[(Y_3 - Y_2) - (Y_2 - Y_1)] = [(Y_4 - Y_3) - (Y_3 - Y_2)] = \dots = [(Y_n - Y_{n-1}) - (Y_{n-1} - Y_{n-2})]$$

- If an exponential trend model provides a perfect fit to a time series, then the percentage differences between consecutive values are constant. Thus,

$$\frac{Y_2 - Y_1}{Y_1} \times 100\% = \frac{Y_3 - Y_2}{Y_2} \times 100\% = \dots = \frac{Y_n - Y_{n-1}}{Y_{n-1}} \times 100\%$$

Although you should not expect a perfectly fitting model for any particular set of time-series data, you can consider the first differences, second differences, and percentage differences as guides in choosing an appropriate model. Examples 16.2, 16.3, and 16.4 illustrate linear, quadratic, and exponential trend models that have perfect (or nearly perfect) fits to their respective data sets.

EXAMPLE 16.2

A Linear Trend Model with a Perfect Fit

The following time series represents the number of customers per year (in thousands) at a branch of a fast-food chain:

	Year									
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Customers Y	200	205	210	215	220	225	230	235	240	245

Using first differences, show that the linear trend model provides a perfect fit to these data.

SOLUTION The following table shows the solution:

	Year									
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Customers Y	200	205	210	215	220	225	230	235	240	245
First differences		5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0

The differences between consecutive values in the series are the same throughout. Thus, the number of customers at the branch of the fast-food chain shows a linear growth pattern.

EXAMPLE 16.3**A Quadratic Trend Model with a Perfect Fit**

The following time series represents the number of customers per year (in thousands) at another branch of a fast-food chain:

	Year									
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Customers Y	200	201	203.5	207.5	213	220	228.5	238.5	250	263

Using second differences, show that the quadratic trend model provides a perfect fit to these data.

SOLUTION The following table shows the solution:

	Year									
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Customers Y	200	201	203.5	207.5	213	220	228.5	238.5	250	263
First differences			1.0	2.5	4.0	5.5	7.0	8.5	10.0	11.5
Second differences				1.5	1.5	1.5	1.5	1.5	1.5	1.5

The second differences between consecutive pairs of values in the series are the same throughout. Thus, branch of the fast-food chain shows a quadratic growth pattern. Its rate of growth is accelerating over time.

EXAMPLE 16.4**An Exponential Trend Model with an Almost Perfect Fit**

The following time series represents the number of customers per year (in thousands) for another branch of the fast-food chain:

	Year									
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Customers Y	200	206	212.18	218.55	225.11	231.86	238.82	245.98	253.36	260.96

Using percentage differences, show that the exponential trend model provides almost a perfect fit to these data.

SOLUTION The following table shows the solution:

	Year									
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Customers Y	200	206	212.18	218.55	225.11	231.86	238.82	245.98	253.36	260.96
Percentage differences		3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0

The percentage differences between consecutive values in the series are approximately the same throughout. Thus, this branch of the fast-food chain shows an exponential growth pattern. Its rate of growth is approximately 3% per year.

Figure 16.11 shows a worksheet that compares the first, second, and percentage differences for the revenues data at The Coca-Cola Company. Neither the first differences, second differences, nor percentage differences are constant across the series. Therefore, other models (including those considered in Section 16.5) may be more appropriate.

FIGURE 16.11

Excel worksheet that compares first, second, and percentage differences in revenues (in \$billions) for The Coca-Cola Company

	A	B	C	D	E
1	Year	Revenues	First Difference	Second Difference	Percentage Difference
2	1996	18.5	#N/A	#N/A	#N/A
3	1997	18.9	0.4	#N/A	2.16%
4	1998	18.8	-0.1	-0.5	-0.53%
5	1999	19.8	1.0	1.1	5.32%
6	2000	20.5	0.7	-0.3	3.54%
7	2001	20.1	-0.4	-1.1	-1.95%
8	2002	19.6	-0.5	-0.1	-2.49%
9	2003	21.0	1.4	1.9	7.14%
10	2004	21.9	0.9	-0.5	4.29%
11	2005	23.1	1.2	0.3	5.48%
12	2006	24.1	1.0	-0.2	4.33%
13	2007	28.9	4.8	3.8	19.92%
14	2008	31.9	3.0	-1.8	10.38%
15	2009	31.0	-0.9	-3.9	-2.82%
16	2010	35.1	4.1	5.0	13.23%
17	2011	46.5	11.4	7.3	32.48%
18	2012	48.0	1.5	-9.9	3.23%

Problems for Section 16.4

LEARNING THE BASICS

16.9 If you are using the method of least squares for fitting trends in an annual time series containing 25 consecutive yearly values,

- what coded value do you assign to X for the first year in the series?
- what coded value do you assign to X for the fifth year in the series?
- what coded value do you assign to X for the most recent recorded year in the series?
- what coded value do you assign to X if you want to project the trend and make a forecast five years beyond the last observed value?

16.10 The linear trend forecasting equation for an annual time series containing 22 values (from 1992 to 2013) on total revenues (in \$millions) is

$$\hat{Y}_i = 4.0 + 1.5X_i$$

- Interpret the Y intercept, b_0 .
- Interpret the slope, b_1 .
- What is the fitted trend value for the fifth year?
- What is the fitted trend value for the most recent year?
- What is the projected trend forecast three years after the last value?

16.11 The linear trend forecasting equation for an annual time series containing 42 values (from 1972 to 2013) on net sales (in \$billions) is

$$\hat{Y}_i = 1.2 + 0.5X_i$$

- Interpret the Y intercept, b_0 .
- Interpret the slope, b_1 .
- What is the fitted trend value for the tenth year?
- What is the fitted trend value for the most recent year?
- What is the projected trend forecast two years after the last value?

APPLYING THE CONCEPTS



16.12 Bed Bath & Beyond is a nationwide chain of retail stores that sell a wide assortment of merchandise, including domestic merchandise and home furnishings, as well as food, giftware, and health and beauty care items. The number of stores open at the end of the fiscal year from 1997 to 2013 is stored in **Bed & Bath** and shown in right column.

Year	Stores Opened	Year	Stores Opened
1997	108	2006	809
1998	141	2007	888
1999	186	2008	971
2000	241	2009	1,037
2001	311	2010	1,100
2002	396	2011	1,139
2003	519	2012	1,173
2004	629	2013	1,471
2005	721		

Source: Data extracted from *Bed Bath & Beyond Annual Report*, 2013.

- Plot the data.
- Compute a linear trend forecasting equation and plot the results.
- Compute a quadratic trend forecasting equation and plot the results.
- Compute an exponential trend forecasting equation and plot the results.
- Using the forecasting equations in (b) through (d), what are your annual forecasts of the number of stores open for 2014 and 2015?
- How can you explain the differences in the three forecasts in (e)? What forecast do you think you should use? Why?

16.13 Gross domestic product (GDP) is a major indicator of a nation's overall economic activity. It consists of personal consumption expenditures, gross domestic investment, net exports of goods and services, and government consumption expenditures. The file **GDP** contains the GDP (in billions of current dollars) for the United States from 1980 to 2012. (Data extracted from Bureau of Economic Analysis, U.S. Department of Commerce, www.bea.gov.)

- Plot the data.
- Compute a linear trend forecasting equation and plot the trend line.
- What are your forecasts for 2013 and 2014?
- What conclusions can you reach concerning the trend in GDP?

16.14 The data in **FedReceipt** represent federal receipts from 1978 through 2012, in billions of current dollars, from individual and corporate income tax, social insurance, excise tax, estate and gift tax, customs duties, and federal reserve deposits. (Data extracted from "Historical Federal Receipt and Outlay Summary," Tax Policy Center, bit.ly/7dGCmz)

- Plot the series of data.
- Compute a linear trend forecasting equation and plot the trend line.
- What are your forecasts of the federal receipts for 2013 and 2014?
- What conclusions can you reach concerning the trend in federal receipts?

16.15 The file **ComputerSales** contains the U.S. total computer and software sales (in \$millions) from 1992 through 2012.

- Plot the data.
- Compute a linear trend forecasting equation and plot the trend line.
- Compute a quadratic trend forecasting equation and plot the results.
- Compute an exponential trend forecasting equation and plot the results.
- Which model is the most appropriate?
- Using the most appropriate model, forecast U.S. total computer and software sales, in millions, for 2013.

16.16 The data shown in the following table and stored in **Solar Power** represent the yearly amount of solar power generated by utilities (in millions of kWh) in the United States from 2002 through 2012:

Year	Solar Power Generated		Year	Solar Power Generated	
	(millions of kWh)			(millions of kWh)	
2002	555		2008	864	
2003	534		2009	891	
2004	575		2010	1,212	
2005	550		2011	1,814	
2006	508		2012	4,432	
2007	612				

Source: Data extracted from en.wikipedia.org/wiki/Solar_power_in_the_United_States.

- Plot the data.
- Compute a linear trend forecasting equation and plot the trend line.
- Compute a quadratic trend forecasting equation and plot the results.
- Compute an exponential trend forecasting equation and plot the results.
- Using the models in (b) through (d), what are your annual trend forecasts of the yearly amount of solar power generated by utilities (in millions of kWh) in the United States in 2013 and 2014?

16.17 The file **CarProduction** contains the number of passenger cars produced in the U.S. from 1999 to 2012. (Data extracted from www.statista.com.)

- Plot the data.
- Compute a linear trend forecasting equation and plot the trend line.
- Compute a quadratic trend forecasting equation and plot the results.
- Compute an exponential trend forecasting equation and plot the results.
- Which model is the most appropriate?
- Using the most appropriate model, forecast the U.S. car production for 2013.

16.18 The average salary of Major League Baseball players on opening day from 2000 to 2013 is stored in **BBSalaries** and shown below.

Year	Salary (\$millions)	Year	Salary (\$millions)
2000	1.99	2007	2.92
2001	2.29	2008	3.13
2002	2.38	2009	3.26
2003	2.58	2010	3.27
2004	2.49	2011	3.32
2005	2.63	2012	3.38
2006	2.83	2013	4.25

Source: Data extracted from "Baseball Salaries," *USA Today*, April 6, 2009, p. 6C; and mlb.com.

- Plot the data.
- Compute a linear trend forecasting equation and plot the trend line.
- Compute a quadratic trend forecasting equation and plot the results.
- Compute an exponential trend forecasting equation and plot the results.
- Which model is the most appropriate?
- Using the most appropriate model, forecast the average salary for 2014.

16.19 The file **Silver** contains the following prices in London for an ounce of silver (in US\$) on the last day of the year from 1999 to 2012:

Year	Price (US\$/ounce)	Year	Price (US\$/ounce)
1999	5.330	2006	12.900
2000	4.570	2007	14.760
2001	4.520	2008	10.790
2002	4.670	2009	16.990
2003	5.965	2010	30.630
2004	6.815	2011	28.180
2005	8.830	2012	29.950

Source: Data extracted from bit.ly/1afifi.

- Plot the data.
- Compute a linear trend forecasting equation and plot the trend line.
- Compute a quadratic trend forecasting equation and plot the results.
- Compute an exponential trend forecasting equation and plot the results.
- Which model is the most appropriate?
- Using the most appropriate model, forecast the price of silver at the end of 2013.

16.20 The data in **CPI-U** reflect the annual values of the consumer price index (CPI) in the United States over the 48-year period 1965 through 2012, using 1982 through 1984 as the base period. This index measures the average change in prices over time in a fixed

“market basket” of goods and services purchased by all urban consumers, including urban wage earners (i.e., clerical, professional, managerial, and technical workers; self-employed individuals; and short-term workers), unemployed individuals, and retirees. (Data extracted from Bureau of Labor Statistics, U.S. Department of Labor, www.bls.gov.)

- Plot the data.
- Describe the movement in this time series over the 48-year period.
- Compute a linear trend forecasting equation and plot the trend line.
- Compute a quadratic trend forecasting equation and plot the results.
- Compute an exponential trend forecasting equation and plot the results.
- Which model is the most appropriate?
- Using the most appropriate model, forecast the CPI for 2013 and 2014.

16.21 Although you should not expect a perfectly fitting model for any time-series data, you can consider the first differences, second differences, and percentage differences for a given series as guides in choosing an appropriate model.

Year	Series I	Series II	Series III
2001	10.0	30.0	60.0
2002	15.1	33.1	67.9
2003	24.0	36.4	76.1
2004	36.7	39.9	84.0
2005	53.8	43.9	92.2
2006	74.8	48.2	100.0
2007	100.0	53.2	108.0
2008	129.2	58.2	115.8
2009	162.4	64.5	124.1
2010	199.0	70.7	132.0
2011	239.3	77.1	140.0
2012	283.5	83.9	147.8

For this problem, use each of the time series presented in the table in the left column and stored in **Tsmodel1**:

- Determine the most appropriate model.
- Compute the forecasting equation.
- Forecast the value for 2013.

16.22 A time-series plot often helps you determine the appropriate model to use. For this problem, use each of the time series presented in the following table and stored in **TsModel2**:

Year	Series I	Series II
2001	100.0	100.0
2002	115.2	115.2
2003	130.1	131.7
2004	144.9	150.8
2005	160.0	174.1
2006	175.0	200.0
2007	189.8	230.8
2008	204.9	266.1
2009	219.8	305.5
2010	235.0	351.8
2011	249.8	403.0
2012	264.9	469.2

- Plot the observed data (Y) over time (X) and plot the logarithm of the observed data ($\log Y$) over time (X) to determine whether a linear trend model or an exponential trend model is more appropriate. (Hint: If the plot of $\log Y$ versus X appears to be linear, an exponential trend model provides an appropriate fit.)
- Compute the appropriate forecasting equation.
- Forecast the value for 2013.

16.5 Autoregressive Modeling for Trend Fitting and Forecasting

Frequently, the values of a time series at particular points in time are highly correlated with the values that precede and succeed them. This type of correlation is called **autocorrelation**. When the autocorrelation exists between values that are in consecutive periods in a time series, the time series displays **first-order autocorrelation**. When the autocorrelation exists between values that are two periods apart, the time series displays **second-order autocorrelation**. For the general case in which the autocorrelation exists between values that are p periods apart, the time series displays **p th-order autocorrelation**.

Autoregressive modeling is a technique used to forecast time series that display autocorrelation.² This type of modeling uses a set of *lagged predictor variables* to overcome the problems that autocorrelation causes with other models. A **lagged predictor variable** takes its value from the value of predictor variable for another time period. For the general case of p th-order autocorrelation, you create a set of p lagged predictor variables such that the first lagged predictor variable takes its value from the value of a predictor variable that is one time period away, the *lag*; that the second lagged predictor variable takes its value from the value of a predictor variable that is two time periods away; and so on until the last, or p th, lagged predictor variable that takes its value from the value of a predictor variable that is p time periods away.

Equation (16.8) defines the **p th-order autoregressive model**. In the equation, A_0, A_1, \dots, A_p represent the parameters and a_0, a_1, \dots, a_p represent the corresponding regression coefficients. This is similar to the multiple regression model, Equation (14.1) on page 545,

²The exponential smoothing model described in Section 16.3 and the autoregressive models described in this section are special cases of autoregressive integrated moving average (ARIMA) models developed by Box and Jenkins (see reference 2).

in which $\beta_0, \beta_1, \dots, \beta_k$, represent the regression parameters and b_0, b_1, \dots, b_k represent the corresponding regression coefficients.

pTH-ORDER AUTOREGRESSIVE MODELS

$$Y_i = A_0 + A_1 Y_{i-1} + A_2 Y_{i-2} + \cdots + A_p Y_{i-p} + \delta_i \quad (16.8)$$

where

Y_i = observed value of the series at time i

Y_{i-1} = observed value of the series at time $i - 1$

Y_{i-2} = observed value of the series at time $i - 2$

Y_{i-p} = observed value of the series at time $i - p$

p = number of autoregression parameters (not including a Y intercept) to be estimated from least-squares regression analysis

$A_0, A_1, A_2, \dots, A_p$ = autoregression parameters to be estimated from least-squares regression analysis

δ_i = a nonautocorrelated random error component (with mean = 0 and constant variance)



Student Tip

δ is the Greek letter delta

Equations (16.9) and (16.10) define two specific autoregressive models. Equation (16.9) defines the **first-order autoregressive model** and is similar in form to the simple linear regression model, Equation (13.1) on page 493. Equation (16.10) defines the **second-order autoregressive model** and is similar to the multiple regression model with two independent variables, Equation (14.2) on page 545.

FIRST-ORDER AUTOREGRESSIVE MODEL

$$Y_i = A_0 + A_1 Y_{i-1} + \delta_i \quad (16.9)$$

SECOND-ORDER AUTOREGRESSIVE MODEL

$$Y_i = A_0 + A_1 Y_{i-1} + A_2 Y_{i-2} + \delta_i \quad (16.10)$$

Selecting an Appropriate Autoregressive Model

Selecting an appropriate autoregressive model can be complicated. You must weigh the advantages of using a simpler model against the concern of not taking into account important autocorrelation in the data. You also must be concerned with selecting a higher-order model that requires estimates of numerous parameters, some of which may be unnecessary, especially if n , the number of values in the series, is small. The reason for this concern is that when computing an estimate of A_p , you lose p out of the n data values when comparing each data value with the data value p periods earlier. Examples 16.5 and 16.6 illustrate this loss of data values.

EXAMPLE 16.5

Consider the following series of $n = 7$ consecutive annual values:

Comparison Schema for a First-Order Autoregressive Model

Series	Year						
	1	2	3	4	5	6	7
	31	34	37	35	36	43	40

Show the comparisons needed for a first-order autoregressive model.

SOLUTION

Year i	First-Order Autoregressive Model (Lag1: Y_i versus Y_{i-1})
1	$31 \leftrightarrow \dots$
2	$34 \leftrightarrow 31$
3	$37 \leftrightarrow 34$
4	$35 \leftrightarrow 37$
5	$36 \leftrightarrow 35$
6	$43 \leftrightarrow 36$
7	$40 \leftrightarrow 43$

Because Y_1 is the first value and there is no value prior to it, Y_1 is not used in the regression analysis. Therefore, the first-order autoregressive model would be based on six pairs of values.

EXAMPLE 16.6

Consider the following series of $n = 7$ consecutive annual values:

Comparison Schema for a Second-Order Autoregressive Model

Series	Year						
	1	2	3	4	5	6	7
	31	34	37	35	36	43	40

Show the comparisons needed for a second-order autoregressive model.

SOLUTION

Year i	Second-Order Autoregressive Model (Lag2: Y_i vs. Y_{i-1} and Y_{i-2})
1	$31 \leftrightarrow \dots$ and $31 \leftrightarrow \dots$
2	$34 \leftrightarrow 31$ and $34 \leftrightarrow \dots$
3	$37 \leftrightarrow 34$ and $37 \leftrightarrow 31$
4	$35 \leftrightarrow 37$ and $35 \leftrightarrow 34$
5	$36 \leftrightarrow 35$ and $36 \leftrightarrow 37$
6	$43 \leftrightarrow 36$ and $43 \leftrightarrow 35$
7	$40 \leftrightarrow 43$ and $40 \leftrightarrow 36$

Because no value is recorded prior to Y_1 , the first two comparisons, each of which requires a value prior to Y_1 , cannot be used when performing regression analysis. Therefore, the second-order autoregressive model would be based on five pairs of values.

Determining the Appropriateness of a Selected Model

After selecting a model and using the least-squares method to compute the regression coefficients, you need to determine the appropriateness of the model. Either you can select a particular p th-order autoregressive model based on previous experiences with similar data or start with a model that contains several autoregressive parameters and then eliminate the higher-order parameters that do not significantly contribute to the model. In this latter approach, you use a t test for the significance of A_p , the highest-order autoregressive parameter in the current model under consideration. The null and alternative hypotheses are:

$$\begin{aligned} H_0: A_p &= 0 \\ H_1: A_p &\neq 0 \end{aligned}$$

Equation (16.11) defines the test statistic.

***t* TEST FOR SIGNIFICANCE OF THE HIGHEST-ORDER AUTOREGRESSIVE PARAMETER, A_p**

$$t_{STAT} = \frac{a_p - A_p}{S_{a_p}} \quad (16.11)$$

where

A_p = hypothesized value of the highest-order parameter, A_p , in the autoregressive model

a_p = regression coefficient that estimates the highest-order parameter, A_p , in the autoregressive model

S_{a_p} = standard deviation of a_p

The t_{STAT} test statistic follows a t distribution with $n - 2p - 1$ degrees of freedom.

In addition to the degrees of freedom lost for each of the p population parameters you are estimating, p additional degrees of freedom are lost because there are p fewer comparisons to be made from the original n values in the time series.

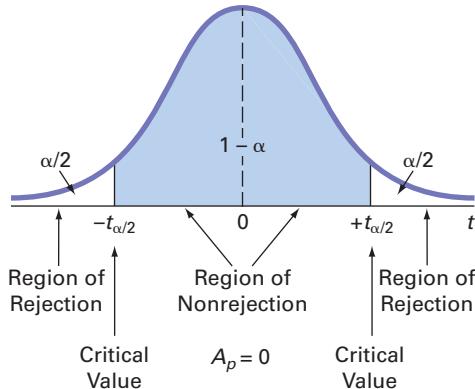
For a given level of significance, α , you reject the null hypothesis if the t_{STAT} test statistic is greater than the upper-tail critical value from the t distribution or if the t_{STAT} test statistic is less than the lower-tail critical value from the t distribution. Thus, the decision rule is

Reject H_0 if $t_{STAT} < -t_{\alpha/2}$ or if $t_{STAT} > t_{\alpha/2}$;
otherwise, do not reject H_0 .

Figure 16.12 illustrates the decision rule and regions of rejection and nonrejection.

FIGURE 16.12

Rejection regions for a two-tail test for the significance of the highest-order autoregressive parameter A_p



If you do not reject the null hypothesis that $A_p = 0$, you conclude that the selected model contains too many estimated autoregressive parameters. You then discard the highest-order term and develop an autoregressive model of order $p - 1$, using the least-squares method. You then repeat the test of the hypothesis that the new highest-order parameter is 0. This testing and modeling continues until you reject H_0 . When this occurs, you can conclude that the remaining highest-order parameter is significant, and you can use that model for forecasting purposes.

Equation (16.12) defines the fitted p th-order autoregressive equation.

FITTED p TH-ORDER AUTOREGRESSIVE EQUATION

$$\hat{Y}_i = a_0 + a_1 Y_{i-1} + a_2 Y_{i-2} + \cdots + a_p Y_{i-p} \quad (16.12)$$

where

- \hat{Y}_i = fitted values of the series at time i
- Y_{i-1} = observed value of the series at time $i - 1$
- Y_{i-2} = observed value of the series at time $i - 2$
- Y_{i-p} = observed value of the series at time $i - p$
- p = number of autoregression parameters (not including a Y intercept) to be estimated from least-squares regression analysis
- $a_0, a_1, a_2, \dots, a_p$ = regression coefficients

You use Equation (16.13) to forecast j years into the future from the current n th time period.

p TH-ORDER AUTOREGRESSIVE FORECASTING EQUATION

$$\hat{Y}_{n+j} = a_0 + a_1 \hat{Y}_{n+j-1} + a_2 \hat{Y}_{n+j-2} + \cdots + a_p \hat{Y}_{n+j-p} \quad (16.13)$$

where

- $a_0, a_1, a_2, \dots, a_p$ = regression coefficients that estimate the parameters
- p = number of autoregression parameters (not including a Y intercept) to be estimated from least-squares regression analysis
- j = number of years into the future
- \hat{Y}_{n+j-p} = forecast of Y_{n+j-p} from the current year for $j - p > 0$
- \hat{Y}_{n+j-p} = observed value for Y_{n+j-p} for $j - p \leq 0$

Thus, to make forecasts j years into the future, using a third-order autoregressive model, you need only the most recent $p = 3$ values (Y_n , Y_{n-1} , and Y_{n-2}) and the regression estimates a_0 , a_1 , a_2 , and a_3 .

To forecast one year ahead, Equation (16.13) becomes

$$\hat{Y}_{n+1} = a_0 + a_1 Y_n + a_2 Y_{n-1} + a_3 Y_{n-2}$$

To forecast two years ahead, Equation (16.13) becomes

$$\hat{Y}_{n+2} = a_0 + a_1 \hat{Y}_{n+1} + a_2 Y_n + a_3 Y_{n-1}$$

To forecast three years ahead, Equation (16.13) becomes

$$\hat{Y}_{n+3} = a_0 + a_1 \hat{Y}_{n+2} + a_2 \hat{Y}_{n+1} + a_3 Y_n$$

and so on.

Autoregressive modeling is a powerful forecasting technique for time series that have autocorrelation. To summarize, you construct an autoregressive model by following these steps:

1. Choose a value for p , the highest-order parameter in the autoregressive model to be evaluated, realizing that the t test for significance is based on $n - 2p - 1$ degrees of freedom.
2. Create a set of p lagged predictor variables. (See Figure 16.13 for an example.)

 **Student Tip**

Remember that in an autoregressive model, the independent variable(s) are equal to the dependent variable lagged by a certain number of time periods.

3. Perform a least-squares analysis of the multiple regression model containing all p lagged predictor variables using Excel or Minitab.
 4. Test for the significance of A_p , the highest-order autoregressive parameter in the model.
 5. If you do not reject the null hypothesis, discard the p th variable and repeat steps 3 and 4. The test for the significance of the new highest-order parameter is based on a t distribution whose degrees of freedom are revised to correspond with the revised number of predictors.
- If you reject the null hypothesis, select the autoregressive model with all p predictors for fitting [see Equation (16.12)] and forecasting [see Equation (16.13)].

To demonstrate the autoregressive modeling approach, return to the time series concerning the revenues for The Coca-Cola Company over the 17-year period 1996 through 2012. Figure 16.13 displays a worksheet that organizes the data for the first-order, second-order, and third-order autoregressive models.

FIGURE 16.13

Minitab worksheet data for developing first-order, second-order, and third-order autoregressive models of the revenues for The Coca-Cola Company (1996–2012)

+	C1	C2	C3	C4	C5
Year	Revenues	Lag1	Lag2	Lag3	
1	1996	18.5	*	*	*
2	1997	18.9	18.5	*	*
3	1998	18.8	18.9	18.5	*
4	1999	19.8	18.8	18.9	18.5
5	2000	20.5	19.8	18.8	18.9
6	2001	20.1	20.5	19.8	18.8
7	2002	19.6	20.1	20.5	19.8
8	2003	21.0	19.6	20.1	20.5
9	2004	21.9	21.0	19.6	20.1
10	2005	23.1	21.9	21.0	19.6
11	2006	24.1	23.1	21.9	21.0
12	2007	28.9	24.1	23.1	21.9
13	2008	31.9	28.9	24.1	23.1
14	2009	31.0	31.9	28.9	24.1
15	2010	35.1	31.0	31.9	28.9
16	2011	46.5	35.1	31.0	31.0
17	2012	48.0	46.5	35.1	31.0

The worksheet contains the lagged predictor variables Lag1, Lag2, and Lag3 in columns C3, C4, and C5. Use all three lagged predictors to fit the third-order autoregressive model. Use only Lag1 and Lag2 to fit the second-order autoregressive model, and use only Lag1 to fit the first-order autoregressive models. Thus, out of $n = 17$ values, $p = 1, 2$, or 3 values out of $n = 17$ are lost in the comparisons needed for developing the first-order, second-order, and third-order autoregressive models.

Selecting an autoregressive model that best fits the annual time series begins with the third-order autoregressive model shown in Figure 16.14.

FIGURE 16.14

Excel and Minitab regression results for a third-order autoregressive model for The Coca-Cola Company revenues

Regression Analysis: Revenues versus Lag1, Lag2, Lag3						
The regression equation is Revenues = -12.1 + 0.797 Lag1 - 0.930 Lag2 + 1.83 Lag3						
Predictor	Coef	SE Coef	T	P		
Constant	-12.072	2.4893	-4.85	0.001		
Lag1	0.7974	0.1733	4.60	0.001		
Lag2	-0.9298	0.3625	-2.57	0.028		
Lag3	1.8322	0.3185	5.75	0.000		
$S = 1.50243 \quad R-Sq = 98.1\% \quad R-Sq(adj) = 97.6\%$						
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	3	1176.98	392.33	173.80	0.000	
Residual Error	10	22.57	2.26			
Total	13	1199.55				
Coefficients Standard Error t Stat P-value Lower 95% Upper 95%						
Intercept	-12.0725	2.4893	-4.8497	0.0007	-17.6191	-6.5259
Lag1	0.7974	0.1733	4.6012	0.0010	0.4112	1.1835
Lag2	-0.9298	0.3625	-2.5651	0.0281	-1.7374	-0.1221
Lag3	1.8322	0.3185	5.7529	0.0002	1.1226	2.5419

From Figure 16.14, the fitted third-order autoregressive equation is

$$\hat{Y}_t = -12.0725 + 0.7974Y_{t-1} - 0.9298Y_{t-2} + 1.8322Y_{t-3}$$

where the first year in the series is 1999.

Next, you test for the significance of A_3 , the highest-order parameter. The highest-order regression coefficient, a_3 , for the fitted third-order autoregressive model is 1.8322, with a standard error of 0.3185.

To test the null hypothesis:

$$H_0: A_3 = 0$$

against the alternative hypothesis:

$$H_1: A_3 \neq 0$$

using Equation (16.11) on page 650 and the worksheet results given in Figure 16.14,

$$t_{STAT} = \frac{a_3 - A_3}{S_{a_3}} = \frac{1.8322 - 0}{0.3185} = 5.7529$$

Using a 0.05 level of significance, the two-tail t test with $14 - 3 - 1 = 10$ degrees of freedom has critical values of ± 2.2281 . Because $t_{STAT} = 5.7529 > 2.2281$ or because the p -value = 0.0002 < 0.05, you reject H_0 . You conclude that the third-order parameter of the autoregressive model is significant and should remain in the model.

The model-building approach has led to the selection of the third-order autoregressive model as the most appropriate for the given data. Using the estimates $a_0 = -12.0725$, $a_1 = 0.7974$, $a_2 = -0.9298$, and $a_3 = 1.8322$, as well as the most recent data value $Y_{16} = 48.0$, the forecasts of revenues from Equation (16.13) on page 651 at The Coca-Cola Company for 2013 and 2014 are

$$\hat{Y}_{n+j} = -12.0725 + 0.7974 \hat{Y}_{n+j-1} - 0.9298 \hat{Y}_{n+j-2} + 1.8322 \hat{Y}_{n+j-3}$$

Therefore, for 2013, one year ($j = 1$) ahead:

$$\begin{aligned}\hat{Y}_{17} &= -12.0725 + 0.7974(48.0) - 0.9298(46.5) + 1.8322(35.1) \\ &= 47.2722 \text{ billions of dollars}\end{aligned}$$

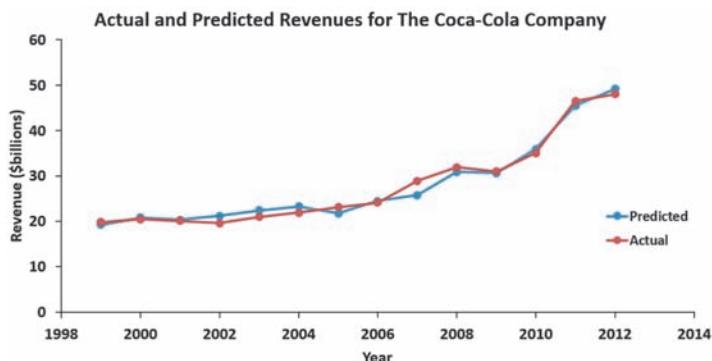
and for 2014, two years ($j = 2$) ahead:

$$\begin{aligned}\hat{Y}_{18} &= -12.0725 + 0.7974(47.7722) - 0.9298(48.0) + 1.8322(46.5) \\ &= 66.19 \text{ billions of dollars}\end{aligned}$$

Figure 16.15 displays the actual and predicted Y values from the third-order autoregressive model.

FIGURE 16.15

Plot of actual and predicted revenues from a third-order autoregressive model for The Coca-Cola Company



Problems for Section 16.5

LEARNING THE BASICS

16.23 You are given an annual time series with 40 consecutive values and asked to fit a fifth-order autoregressive model.

- How many comparisons are lost in developing the autoregressive model?
- How many parameters do you need to estimate?
- Which of the original 40 values do you need for forecasting?
- State the fifth-order autoregressive model.
- Write an equation to indicate how you would forecast j years into the future.

16.24 A third-order autoregressive model is fitted to an annual time series with 17 values and has the following estimated parameters and standard errors:

$$\begin{aligned} a_0 &= 4.50 & a_1 &= 1.80 & a_2 &= 0.80 & a_3 &= 0.24 \\ S_{a_1} &= 0.50 & S_{a_2} &= 0.30 & S_{a_3} &= 0.10 \end{aligned}$$

At the 0.05 level of significance, test the appropriateness of the fitted model.

16.25 Refer to Problem 16.24. The three most recent values are

$$Y_{15} = 23 \quad Y_{16} = 28 \quad Y_{17} = 34$$

Forecast the values for the next year and the following year.

16.26 Refer to Problem 16.24. Suppose, when testing for the appropriateness of the fitted model, the standard errors are

$$S_{a_1} = 0.45 \quad S_{a_2} = 0.35 \quad S_{a_3} = 0.15$$

- What conclusions can you reach?
- Discuss how to proceed if forecasting is still your main objective.

APPLYING THE CONCEPTS

16.27 Using the data for Problem 16.15 on page 646 that represent U.S. total computer and software sales (in \$millions) from 1992 through 2012 (stored in **ComputerSales**),

- fit a third-order autoregressive model to the total sales and test for the significance of the third-order autoregressive parameter. (Use $\alpha = 0.05$.)
- if necessary, fit a second-order autoregressive model to the total sales and test for the significance of the second-order autoregressive parameter. (Use $\alpha = 0.05$.)
- if necessary, fit a first-order autoregressive model to the total sales and test for the significance of the first-order autoregressive parameter. (Use $\alpha = 0.05$.)
- if appropriate, forecast the total sales in 2013.

 **16.28** Using the data for Problem 16.12 on page 645 concerning the number of stores open for Bed Bath & Beyond from 1997 through 2013 (stored in **Bed & Bath**),

- fit a third-order autoregressive model to the number of stores and test for the significance of the third-order autoregressive parameter. (Use $\alpha = 0.05$.)
- if necessary, fit a second-order autoregressive model to the number of stores and test for the significance of the second-order autoregressive parameter. (Use $\alpha = 0.05$.)
- if necessary, fit a first-order autoregressive model to the number of stores and test for the significance of the first-order autoregressive parameter. (Use $\alpha = 0.05$.)
- if appropriate, forecast the number of stores open in 2014 and 2015.

16.29 Using the data for Problem 16.17 on page 646 concerning the number of passenger cars produced in the United States from 1999 to 2012 (stored in **CarProduction**),

- fit a third-order autoregressive model to the number of passenger cars produced in the United States and test for the significance of the third-order autoregressive parameter. (Use $\alpha = 0.05$.)
- if necessary, fit a second-order autoregressive model to the number of passenger cars produced in the United States and test for the significance of the second-order autoregressive parameter. (Use $\alpha = 0.05$.)
- if necessary, fit a first-order autoregressive model to the number of passenger cars produced in the United States and test for the significance of the first-order autoregressive parameter. (Use $\alpha = 0.05$.)
- forecast the U.S. car production for 2013.

16.30 Using the average baseball salary from 2000 through 2013 data for Problem 16.18 on page 646 (stored in **BBSalaries**),

- fit a third-order autoregressive model to the average baseball salary and test for the significance of the third-order autoregressive parameter. (Use $\alpha = 0.05$.)
- if necessary, fit a second-order autoregressive model to the average baseball salary and test for the significance of the second-order autoregressive parameter. (Use $\alpha = 0.05$.)
- if necessary, fit a first-order autoregressive model to the average baseball salary and test for the significance of the first-order autoregressive parameter. (Use $\alpha = 0.05$.)
- forecast the average baseball salary for 2014.

16.31 Using the yearly amount of solar power generated by utilities (in millions of kWh) in the United States from 2002 through 2012 data for Problem 16.16 on page 646 (stored in **SolarPower**),

- fit a third-order autoregressive model to the amount of solar power installed and test for the significance of the third-order autoregressive parameter. (Use $\alpha = 0.05$.)
- if necessary, fit a second-order autoregressive model to the amount of solar power installed and test for the significance of the second-order autoregressive parameter. (Use $\alpha = 0.05$.)
- if necessary, fit a first-order autoregressive model to the amount of solar power installed and test for the significance of the first-order autoregressive parameter. (Use $\alpha = 0.05$.)
- forecast the yearly amount of solar power generated by utilities (in millions of kWh) in the United States in 2013 and 2014.

16.6 Choosing an Appropriate Forecasting Model

In Sections 16.4 and 16.5, you studied six time-series methods for forecasting: the linear trend model, the quadratic trend model, and the exponential trend model in Section 16.4; and the first-order, second-order, and p th-order autoregressive models in Section 16.5. Is there a *best* model? Among these models, which one should you select for forecasting? The following guidelines are provided for determining the adequacy of a particular forecasting model. These guidelines are based on a judgment of how well the model fits the data and assume that you can use past data to predict future values of the time series:

- Perform a residual analysis.
- Measure the magnitude of the residuals through squared differences.
- Measure the magnitude of the residuals through absolute differences.
- Use the principle of parsimony.

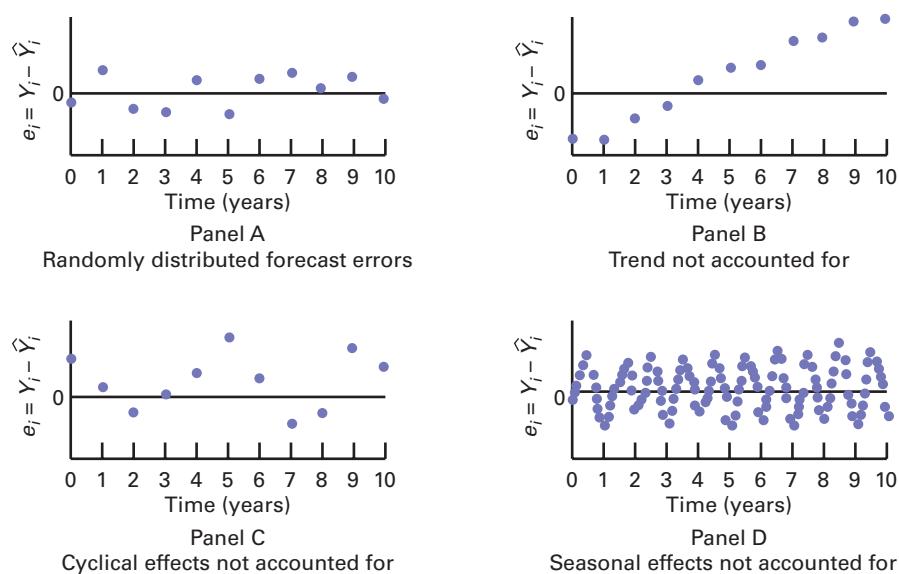
A discussion of these guidelines follows.

Performing a Residual Analysis

Recall from Sections 13.5 and 14.3 that residuals are the differences between observed and predicted values. After fitting a particular model to a time series, you plot the residuals over the n time periods. As shown in Figure 16.16 Panel A, if the particular model fits adequately, the residuals represent the irregular component of the time series. Therefore, they should be randomly distributed throughout the series. However, as illustrated in the three remaining panels of Figure 16.16, if the particular model does not fit adequately, the residuals may show a systematic pattern, such as a failure to account for trend (Panel B), a failure to account for cyclical variation (Panel C), or, with monthly or quarterly data, a failure to account for seasonal variation (Panel D).

FIGURE 16.16

Residual analysis for studying patterns of errors in regression models



Measuring the Magnitude of the Residuals Through Squared or Absolute Differences

If, after performing a residual analysis, you still believe that two or more models appear to fit the data adequately, you can use additional methods for model selection. Numerous measures based on the residuals are available (see references 1 and 4).

In regression analysis (see Section 13.3), you have already used the standard error of the estimate S_{YX} as a measure of variation around the predicted values. For a particular model, this measure is based on the sum of squared differences between the actual and predicted values in a time series. If a model fits the time-series data perfectly, then the standard error of the

estimate is zero. If a model fits the time-series data poorly, then S_{YX} is large. Thus, when comparing the adequacy of two or more forecasting models, you can select the model with the smallest S_{YX} as most appropriate.

However, a major drawback to using S_{YX} when comparing forecasting models is that whenever there is a large difference between even a single Y_i and \hat{Y}_i , the value of S_{YX} becomes overly inflated because the differences between Y_i and \hat{Y}_i are squared. For this reason, many statisticians prefer the **mean absolute deviation (MAD)**. Equation (16.14) defines the *MAD* as the mean of the absolute differences between the actual and predicted values in a time series.

MEAN ABSOLUTE DEVIATION

$$MAD = \frac{\sum_{i=1}^n |Y_i - \hat{Y}_i|}{n} \quad (16.14)$$

If a model fits the time-series data perfectly, the *MAD* is zero. If a model fits the time-series data poorly, the *MAD* is large. When comparing two or more forecasting models, you can select the one with the smallest *MAD* as the most appropriate model.

Using the Principle of Parsimony

If, after performing a residual analysis and comparing the S_{YX} and *MAD* measures, you still believe that two or more models appear to adequately fit the data, you can use the principle of parsimony for model selection. As first explained in Section 15.4, **parsimony** guides you to select the regression model with the fewest independent variables that can predict the dependent variable adequately. In general, the principle of parsimony guides you to select the least complex regression model. Among the six forecasting models studied in this chapter, most statisticians consider the least-squares linear and quadratic models and the first-order autoregressive model as simpler than the second- and p th-order autoregressive models and the least-squares exponential model.

A Comparison of Four Forecasting Methods

To illustrate the model selection process, you can compare four of the forecasting models used in Sections 16.4 and 16.5: the linear model, the quadratic model, the exponential model, and the third-order autoregressive model. Figure 16.17 shows the residual plots for the four models for The Coca-Cola Company revenues. In reaching conclusions from these residual plots, you must use caution because there are only 17 values for the linear model, the quadratic model, and the exponential model and only 14 values for the third-order autoregressive model.

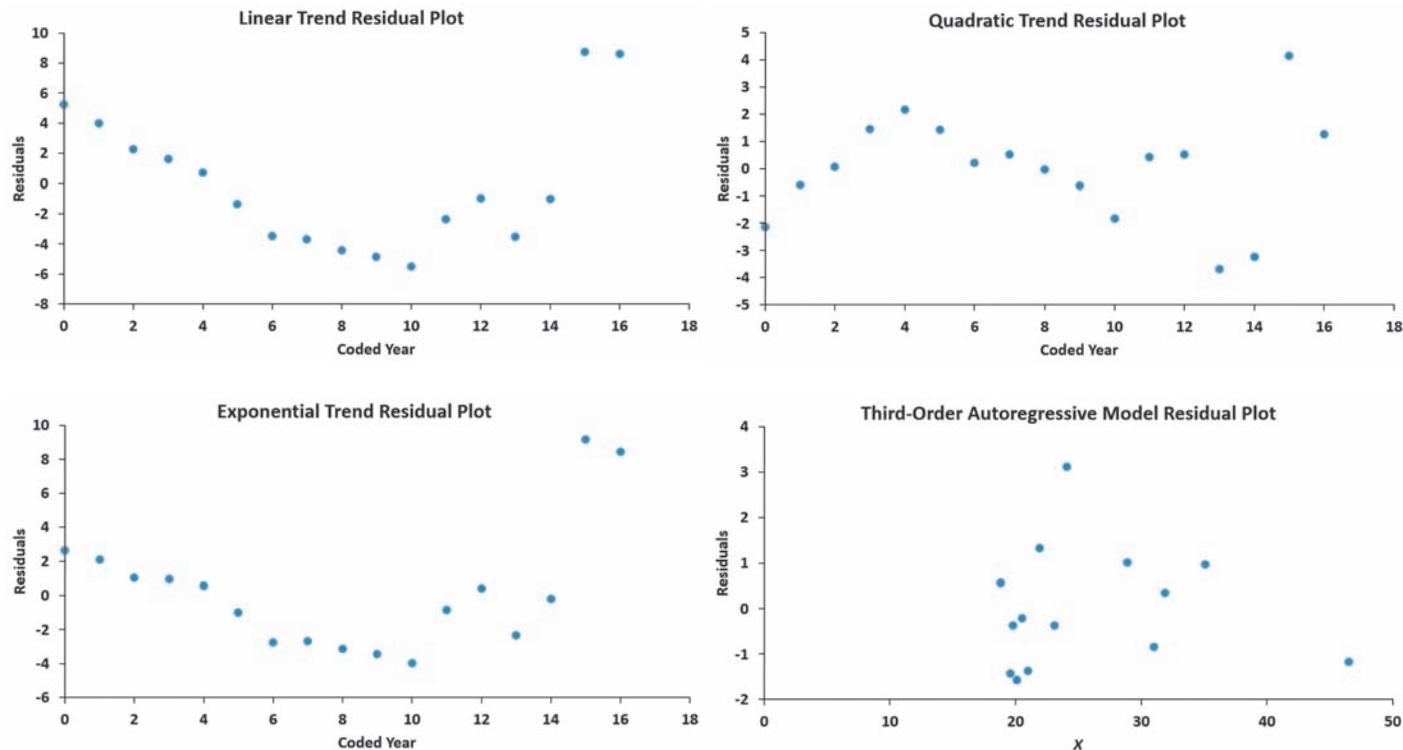
In Figure 16.17, observe that the residuals in the linear model, quadratic model, and exponential model are positive for the early years, negative for the intermediate years, and positive again for the latest years. For the autoregressive model, the residuals do not exhibit any systematic pattern.

To summarize, on the basis of the residual analysis of all four forecasting models, it appears that the third-order autoregressive model is the most appropriate, and the linear, quadratic, and exponential models are not appropriate. For further verification, you can compare the magnitude of the residuals in the four models. Figure 16.18 shows the actual values (Y_i) along with the predicted values \hat{Y}_i , the residuals (e_i), the error sum of squares (SSE), the standard error of the estimate (S_{YX}), and the mean absolute deviation (MAD) for each of the four models.

For this time series, S_{YX} and *MAD* provide fairly similar results. A comparison of the S_{YX} and *MAD* clearly indicates that the linear model provides the poorest fit followed by the exponential model and then the quadratic model. The third-order autoregressive model provides the best fit. Thus, you should choose the third-order autoregressive model as the best model.

FIGURE 16.17

Residual plots for four forecasting models

**FIGURE 16.18**

Comparison of four forecasting models using SSE , S_{YX} , and MAD

Year	Revenues	Linear		Quadratic		Exponential		Third-Order AR	
		Predicted	Residual	Predicted	Residual	Predicted	Residual	Predicted	Residual
1996	18.5	13.2745	5.2255	20.6249	-2.1249	15.8690	2.6310	#N/A	#N/A
1997	18.9	14.9071	3.9929	19.5011	-0.6011	16.8009	2.0991	#N/A	#N/A
1998	18.8	16.5397	2.2603	18.7448	0.0552	17.7876	1.0124	#N/A	#N/A
1999	19.8	18.1723	1.6277	18.3561	1.4439	18.8322	0.9678	19.2411	0.5589
2000	20.5	19.8049	0.6951	18.3348	2.1652	19.9382	0.5618	20.8643	-0.3643
2001	20.1	21.4375	-1.3375	18.6811	1.4189	21.1091	-1.0091	20.3094	-0.2094
2002	19.6	23.0701	-3.4701	19.3949	0.2051	22.3488	-2.7488	21.1719	-1.5719
2003	21.0	24.7027	-3.7027	20.4762	0.5238	23.6613	-2.6613	22.4276	-1.4276
2004	21.9	26.3353	-4.4353	21.9251	-0.0251	25.0509	-3.1509	23.2760	-1.3760
2005	23.1	27.9679	-4.8679	23.7414	-0.6414	26.5221	-3.4221	21.7758	1.3242
2006	24.1	29.6005	-5.5005	25.9253	-1.8253	28.0797	-3.9797	24.4609	-0.3609
2007	28.9	31.2331	-2.3331	28.4767	0.4233	29.7288	-0.8288	25.7915	3.1085
2008	31.9	32.8657	-0.9657	31.3956	0.5044	31.4747	0.4253	30.8878	1.0122
2009	31.0	34.4983	-3.4983	34.6820	-3.6820	33.3231	-2.3231	30.6491	0.3509
2010	35.1	36.1309	-1.0309	38.3360	-3.2360	35.2801	-0.1801	35.9368	-0.8368
2011	46.5	37.7635	8.7365	42.3575	4.1425	37.3521	9.1479	45.5395	0.9605
2012	48.0	39.3961	8.6039	46.7464	1.2536	39.5457	8.4543	49.1683	-1.1683
		SSE 322.6853		SSE 60.9205		SSE 228.1973		SSE 22.5730	
		S_{YX} 4.6381		S_{YX} 2.0860		S_{YX} 3.3798		S_{YX} 1.5024	
		MAD 3.6638		MAD 1.4277		MAD 2.6826		MAD 1.0450	

After you select a particular forecasting model, you need to continually monitor your forecasts. If large errors between forecasted and actual values occur, the underlying structure of the time series may have changed. Remember that the forecasting methods presented in this chapter assume that the patterns inherent in the past will continue into the future. Large forecasting errors are an indication that this assumption may no longer be true.

Problems for Section 16.6

LEARNING THE BASICS

16.32 The following residuals are from a linear trend model used to forecast sales:

2.0 -0.5 1.5 1.0 0.0 1.0 -3.0 1.5 -4.5 2.0 0.0 -1.0

- Compute S_{YX} and interpret your findings.
- Compute the MAD and interpret your findings.

16.33 Refer to Problem 16.32. Suppose the first residual is 12.0 (instead of 2.0) and the last residual is -11.0 (instead of -1.0).

- Compute S_{YX} and interpret your findings.
- Compute the MAD and interpret your findings.

APPLYING THE CONCEPTS

16.34 Using the yearly amount of solar power generated by utilities (in millions of kWh) in the United States from 2002 through 2012 data for Problem 16.16 on page 646 and Problem 16.31 on page 654 (stored in **SolarPower**),

- perform a residual analysis.
- compute the standard error of the estimate (S_{YX}).
- compute the MAD .
- On the basis of (a) through (c), and the principle of parsimony, which forecasting model would you select? Discuss.

16.35 Using the U.S. total computer and software sales data for Problem 16.15 on page 646 and Problem 16.27 on page 654 (stored in **ComputerSales**),

- perform a residual analysis for each model.
- compute the standard error of the estimate (S_{YX}) for each model.
- compute the MAD for each model.
- On the basis of (a) through (c) and the principle of parsimony, which forecasting model would you select? Discuss.

SELF Test **16.36** Using the number of stores open for Bed Bath & Beyond from 1997 through 2013 data for Problem 16.12 on page 645 and Problem 16.28 on page 654 (stored in **Bed & Bath**),

- perform a residual analysis for each model.
- compute the standard error of the estimate (S_{YX}) for each model.
- compute the MAD for each model.
- On the basis of (a) through (c) and the principle of parsimony, which forecasting model would you select? Discuss.

16.37 Using the number of passenger cars produced in the U.S. from 1999 to 2012 data for Problem 16.17 on page 646 and Problem 16.29 on page 654 (stored in **CarProduction**),

- perform a residual analysis for each model.
- compute the standard error of the estimate (S_{YX}) for each model.
- compute the MAD for each model.
- On the basis of (a) through (c) and the principle of parsimony, which forecasting model would you select? Discuss.

16.38 Using the average baseball salary from 2000 through 2013 data for Problem 16.18 on page 646 and Problem 16.30 on page 654 (stored in **BBSalaries**),

- perform a residual analysis for each model.
- compute the standard error of the estimate (S_{YX}) for each model.
- compute the MAD for each model.
- On the basis of (a) through (c) and the principle of parsimony, which forecasting model would you select? Discuss.

16.39 Refer to the results for Problem 16.13 on page 646 that used the file **GDP**,

- perform a residual analysis.
- compute the standard error of the estimate (S_{YX}).
- compute the MAD .
- On the basis of (a) through (c), are you satisfied with your linear trend forecasts in Problem 16.13? Discuss.

16.7 Time-Series Forecasting of Seasonal Data

So far, this chapter has focused on forecasting annual data. However, many time series are collected quarterly or monthly, and others are collected weekly, daily, and even hourly. When a time series is collected quarterly or monthly, you must consider the impact of seasonal effects. In this section, regression model building is used to forecast monthly or quarterly data.

One of the companies of interest in the Using Statistics scenario is Wal-Mart Stores, Inc. In 2013, according to the company's website, Wal-Mart Stores, Inc., operated more than 10,000 retail units in 27 countries and had revenues that exceeded \$400 billion. Walmart revenues are highly seasonal, and therefore you need to analyze quarterly revenues. The fiscal year for the company ends January 31. Thus, the fourth quarter of 2013 includes November and December 2012 as well as January 2013. Table 16.3 lists the quarterly revenues (in \$billions) from 2008 to 2013 that are stored in **Walmart**. Figure 16.19 displays the time series.

TABLE 16.3

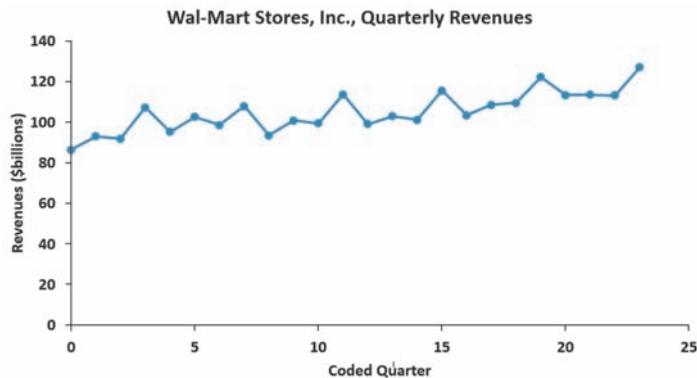
Quarterly Revenues (in \$billions) for Wal-Mart Stores, Inc., 2008–2013

Quarter	Year					
	2008	2009	2010	2011	2012	2013
1	86.4	95.3	93.5	99.1	103.4	113.4
2	93.0	102.7	100.9	103.0	108.6	113.5
3	91.9	98.6	99.4	101.2	109.5	113.2
4	107.3	107.9	113.7	115.6	122.3	127.1

Source: Data extracted from Wal-Mart Stores, Inc., walmartstores.com.

FIGURE 16.19

Plot of quarterly revenues (\$billions) for Wal-Mart Stores, Inc., 2008–2013

**Least-Squares Forecasting with Monthly or Quarterly Data**

To develop a least-squares regression model that includes a seasonal component, the least-squares exponential trend fitting method used in Section 16.4 is combined with dummy variables (see Section 14.6) to model the seasonal component.

Equation (16.15) defines the exponential trend model for quarterly data.

EXPONENTIAL MODEL WITH QUARTERLY DATA

$$Y_i = \beta_0 \beta_1^{X_i} \beta_2^{Q_1} \beta_3^{Q_2} \beta_4^{Q_3} \varepsilon_i \quad (16.15)$$

where

X_i = coded quarterly value, $i = 0, 1, 2, \dots$

Q_1 = 1 if first quarter, 0 if not first quarter

Q_2 = 1 if second quarter, 0 if not second quarter

Q_3 = 1 if third quarter, 0 if not third quarter

β_0 = Y intercept

$(\beta_1 - 1) \times 100\%$ = quarterly compound growth rate (in %)

β_2 = multiplier for first quarter relative to fourth quarter

β_3 = multiplier for second quarter relative to fourth quarter

β_4 = multiplier for third quarter relative to fourth quarter

ε_i = value of the irregular component for time period i

³Alternatively, you can use base e logarithms. For more information on logarithms, see Section A.3 in Appendix A.

The model in Equation (16.15) is not in the form of a linear regression model. To transform this nonlinear model to a linear model, you use a base 10 logarithmic transformation.³ Taking the logarithm of each side of Equation (16.15) results in Equation (16.16).

TRANSFORMED EXPONENTIAL MODEL WITH QUARTERLY DATA

$$\begin{aligned} \log(Y_i) &= \log(\beta_0 \beta_1^{X_i} \beta_2^{Q_1} \beta_3^{Q_2} \beta_4^{Q_3} \varepsilon_i) \\ &= \log(\beta_0) + \log(\beta_1^{X_i}) + \log(\beta_2^{Q_1}) + \log(\beta_3^{Q_2}) + \log(\beta_4^{Q_3}) + \log(\varepsilon_i) \\ &= \log(\beta_0) + X_i \log(\beta_1) + Q_1 \log(\beta_2) + Q_2 \log(\beta_3) + Q_3 \log(\beta_4) + \log(\varepsilon_i) \end{aligned} \quad (16.16)$$

Equation (16.16) is a linear model that you can estimate using least-squares regression. Performing the regression analysis using $\log(Y_i)$ as the dependent variable and X_i , Q_1 , Q_2 , and Q_3 as the independent variables results in Equation (16.17).

EXPONENTIAL GROWTH WITH QUARTERLY DATA FORECASTING EQUATION

$$\log(\hat{Y}_i) = b_0 + b_1 X_i + b_2 Q_1 + b_3 Q_2 + b_4 Q_3 \quad (16.17)$$

where

- b_0 = estimate of $\log(\beta_0)$ and thus $10^{b_0} = \hat{\beta}_0$
- b_1 = estimate of $\log(\beta_1)$ and thus $10^{b_1} = \hat{\beta}_1$
- b_2 = estimate of $\log(\beta_2)$ and thus $10^{b_2} = \hat{\beta}_2$
- b_3 = estimate of $\log(\beta_3)$ and thus $10^{b_3} = \hat{\beta}_3$
- b_4 = estimate of $\log(\beta_4)$ and thus $10^{b_4} = \hat{\beta}_4$

Equation (16.18) is used for monthly data.

EXPONENTIAL MODEL WITH MONTHLY DATA

$$Y_i = \beta_0 \beta_1^{X_i} \beta_2^{M_1} \beta_3^{M_2} \beta_4^{M_3} \beta_5^{M_4} \beta_6^{M_5} \beta_7^{M_6} \beta_8^{M_7} \beta_9^{M_8} \beta_{10}^{M_9} \beta_{11}^{M_{10}} \beta_{12}^{M_{11}} \varepsilon_i \quad (16.18)$$

where

- X_i = coded monthly value, $i = 0, 1, 2, \dots$
- M_1 = 1 if January, 0 if not January
- M_2 = 1 if February, 0 if not February
- M_3 = 1 if March, 0 if not March
- \vdots
- M_{11} = 1 if November, 0 if not November
- β_0 = Y intercept
- $(\beta_1 - 1) \times 100\%$ = monthly compound growth rate (in %)
- β_2 = multiplier for January relative to December
- β_3 = multiplier for February relative to December
- β_4 = multiplier for March relative to December
- \vdots
- β_{12} = multiplier for November relative to December
- ε_i = value of the irregular component for time period i

The model in Equation (16.18) is not in the form of a linear regression model. To transform this nonlinear model into a linear model, you can use a base 10 logarithm transformation. Taking the logarithm of each side of Equation (16.18) results in Equation (16.19).

TRANSFORMED EXPONENTIAL MODEL WITH MONTHLY DATA

$$\begin{aligned} \log(Y_i) &= \log(\beta_0 \beta_1^{X_i} \beta_2^{M_1} \beta_3^{M_2} \beta_4^{M_3} \beta_5^{M_4} \beta_6^{M_5} \beta_7^{M_6} \beta_8^{M_7} \beta_9^{M_8} \beta_{10}^{M_9} \beta_{11}^{M_{10}} \beta_{12}^{M_{11}} \varepsilon_i) \\ &= \log(\beta_0) + X_i \log(\beta_1) + M_1 \log(\beta_2) + M_2 \log(\beta_3) \\ &\quad + M_3 \log(\beta_4) + M_4 \log(\beta_5) + M_5 \log(\beta_6) + M_6 \log(\beta_7) \\ &\quad + M_7 \log(\beta_8) + M_8 \log(\beta_9) + M_9 \log(\beta_{10}) + M_{10} \log(\beta_{11}) \\ &\quad + M_{11} \log(\beta_{12}) + \log(\varepsilon_i) \end{aligned} \quad (16.19)$$

Equation (16.19) is a linear model that you can estimate using the least-squares method. Performing the regression analysis using $\log(Y_i)$ as the dependent variable and X_i, M_1, M_2, \dots , and M_{11} as the independent variables results in Equation (16.20).

EXPONENTIAL GROWTH WITH MONTHLY DATA FORECASTING EQUATION

$$\begin{aligned}\log(\hat{Y}_i) = & b_0 + b_1X_i + b_2M_1 + b_3M_2 + b_4M_3 + b_5M_4 + b_6M_5 + b_7M_6 \\ & + b_8M_7 + b_9M_8 + b_{10}M_9 + b_{11}M_{10} + b_{12}M_{11}\end{aligned}\quad (16.20)$$

where

$$\begin{aligned}b_0 &= \text{estimate of } \log(\beta_0) \text{ and thus } 10^{b_0} = \hat{\beta}_0 \\ b_1 &= \text{estimate of } \log(\beta_1) \text{ and thus } 10^{b_1} = \hat{\beta}_1 \\ b_2 &= \text{estimate of } \log(\beta_2) \text{ and thus } 10^{b_2} = \hat{\beta}_2 \\ b_3 &= \text{estimate of } \log(\beta_3) \text{ and thus } 10^{b_3} = \hat{\beta}_3 \\ &\vdots \\ b_{12} &= \text{estimate of } \log(\beta_{12}) \text{ and thus } 10^{b_{12}} = \hat{\beta}_{12}\end{aligned}$$

Q_1, Q_2 , and Q_3 are the three dummy variables needed to represent the four quarter periods in a quarterly time series. $M_1, M_2, M_3, \dots, M_{11}$ are the 11 dummy variables needed to represent the 12 months in a monthly time series. In building the model, you use $\log(Y_i)$ instead of Y_i values and then find the regression coefficients by taking the antilog of the regression coefficients developed from Equations (16.17) and (16.20).

Although at first glance these regression models look imposing, when fitting or forecasting for any one time period, the values of all or all but one of the dummy variables in the model are equal to zero, and the equations simplify dramatically. In establishing the dummy variables for quarterly time-series data, the fourth quarter is the base period and has a coded value of zero for each dummy variable. With a quarterly time series, Equation (16.17) reduces as follows:

- For any first quarter: $\log(\hat{Y}_i) = b_0 + b_1X_i + b_2$
- For any second quarter: $\log(\hat{Y}_i) = b_0 + b_1X_i + b_3$
- For any third quarter: $\log(\hat{Y}_i) = b_0 + b_1X_i + b_4$
- For any fourth quarter: $\log(\hat{Y}_i) = b_0 + b_1X_i$

When establishing the dummy variables for each month, December serves as the base period and has a coded value of 0 for each dummy variable. For example, with a monthly time series, Equation (16.20) reduces as follows:

- For any January: $\log(\hat{Y}_i) = b_0 + b_1X_i + b_2$
- For any February: $\log(\hat{Y}_i) = b_0 + b_1X_i + b_3$
- ⋮
- For any November: $\log(\hat{Y}_i) = b_0 + b_1X_i + b_{12}$
- For any December: $\log(\hat{Y}_i) = b_0 + b_1X_i$

To demonstrate the process of model building and least-squares forecasting with a quarterly time series, return to the Wal-Mart Stores, Inc., revenue data (in billions of dollars) originally displayed in Table 16.3 on page 658. The data are from the first quarter of 2008 through the last quarter of 2013. Figure 16.20 shows the Excel and Minitab regression results for the quarterly exponential trend model.

From Figure 16.20, the model fits the data very well. The coefficient of determination $r^2 = 0.9461$, the adjusted $r^2 = 0.9348$, and the overall F test results in an F_{STAT} test statistic of 83.391 (p -value = 0.000). At the 0.05 level of significance, each regression coefficient is

FIGURE 16.20

Excel and Minitab regression results for the quarterly revenue data for Wal-Mart Stores, Inc.

A	B	C	D	E	F	G
1 Quarterly Exponential Trend Model for Wal-Mart Stores, Inc., Revenues						
2						
3 Regression Statistics						
4 Multiple R	0.9727					
5 R Square	0.9461					
6 Adjusted R Square	0.9348					
7 Standard Error	0.0105					
8 Observations	24					
9						
10 ANOVA						
11	df	SS	MS	F	Significance F	
12 Regression	4	0.0365	0.0091	83.3910	0.0000	
13 Residual	19	0.0021	0.0001			
14 Total	23	0.0385				
15						
16	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17 Intercept	2.0073	0.0059	340.6493	0.0000	1.9949	2.0196
18 Coded Quarter	0.0042	0.0003	13.5539	0.0000	0.0036	0.0049
19 Q1	-0.0577	0.0061	-9.4406	0.0000	-0.0705	-0.0449
20 Q2	-0.0392	0.0061	-6.4666	0.0000	-0.0520	-0.0265
21 Q3	-0.0492	0.0060	-8.1466	0.0000	-0.0619	-0.0366

Regression Analysis: Log(Revenues) versus Coded Quarter, Q1, Q2, Q3

The regression equation is
 $\text{Log(Revenues)} = 2.01 + 0.00423 \text{ Coded Quarter}$
 $- 0.0577 \text{ Q1} - 0.0392 \text{ Q2} - 0.0492 \text{ Q3}$

Predictor	Coef	SE Coef	T	P
Constant	2.00727	0.00589	340.65	0.000
Coded Quarter	0.0042349	0.0003124	13.55	0.000
Q1	-0.057676	0.006109	-9.44	0.000
Q2	-0.039248	0.006069	-6.47	0.000
Q3	-0.049247	0.006045	-8.15	0.000

S = 0.0104565 R-Sq = 94.6% R-Sq(adj) = 93.5%

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	4	0.0364711	0.0091178	83.39	0.000
Residual Error	19	0.0020774	0.0001093		
Total	23	0.0385485			

highly statistically significant and contributes to the model. The following summary includes the antilogs of all the regression coefficients:

Regression Coefficient	$b_i = \log \hat{\beta}_i$	$\hat{\beta}_i = \text{antilog}(b_i) = 10^{b_i}$
b_0 : Y intercept	2.0073	101.6951
b_1 : coded quarter	0.0042	1.0097
b_2 : first quarter	-0.0577	0.8756
b_3 : second quarter	-0.0392	0.9137
b_4 : third quarter	-0.0492	0.8929

The interpretations for $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, and $\hat{\beta}_4$ are as follows:

- The Y intercept, $\hat{\beta}_0 = 101.6951$ (in \$billions), is the *unadjusted* forecast for quarterly revenues in the first quarter of 2008, the initial quarter in the time series. *Unadjusted* means that the seasonal component is not incorporated in the forecast.
- The value $(\hat{\beta}_1 - 1) \times 100\% = 0.0097$, or 0.97%, is the estimated *quarterly compound growth rate* in revenues, after adjusting for the seasonal component.
- $\hat{\beta}_2 = 0.8756$ is the seasonal multiplier for the first quarter relative to the fourth quarter; it indicates that there is $1 - 0.8756 = 12.44\%$ less revenue for the first quarter than for the fourth quarter.
- $\hat{\beta}_3 = 0.9137$ is the seasonal multiplier for the second quarter relative to the fourth quarter; it indicates that there is $1 - 0.9137 = 8.63\%$ less revenue for the second quarter than for the fourth quarter.
- $\hat{\beta}_4 = 0.8929$ is the seasonal multiplier for the third quarter relative to the fourth quarter; it indicates that there is $1 - 0.8929 = 10.71\%$ less revenue for the third quarter than for the fourth quarter. Thus, the fourth quarter, which includes the holiday shopping season, has the strongest sales.

Using the regression coefficients b_0 , b_1 , b_2 , b_3 , and b_4 , and Equation (16.17) on page 660, you can make forecasts for selected quarters. As an example, to predict revenues for the fourth quarter of 2013 ($X_i = 23$),

$$\begin{aligned}\log(\hat{Y}_i) &= b_0 + b_1 X_i \\ &= 2.0073 + (0.0042)(23) \\ &= 2.1039\end{aligned}$$

Thus,

$$\log(\hat{Y}_i) = 10^{2.1039} = 127.0282$$

The predicted revenue for the fourth quarter of fiscal 2013 is \$127.0282 billion. To make a forecast for a future time period, such as the first quarter of fiscal 2014 ($X_i = 24, Q_1 = 1$),

$$\begin{aligned}\log(\hat{Y}_i) &= b_0 + b_1 X_i + b_2 Q_1 \\ &= 2.0073 + (0.0042)(24) + (-0.0577)(1) \\ &= 2.0504\end{aligned}$$

Thus,

$$\hat{Y}_i = 10^{2.0504} = 112.3052$$

The predicted revenue for the first quarter of fiscal 2014 is \$112.3052 billion.

Problems for Section 16.7

LEARNING THE BASICS

16.40 In forecasting a monthly time series over a five-year period from January 2009 to December 2013, the exponential trend forecasting equation for January is

$$\log \hat{Y}_i = 2.0 + 0.01X_i + 0.10 \text{ (January)}$$

Take the antilog of the appropriate coefficient from this equation and interpret the

- Y intercept, \hat{b}_0 .
- monthly compound growth rate.
- January multiplier.

16.41 In forecasting daily time-series data, how many dummy variables are needed to account for the seasonal component day of the week?

16.42 In forecasting a quarterly time series over the five-year period from the first quarter of 2009 through the fourth quarter of 2013, the exponential trend forecasting equation is given by

$$\log \hat{Y}_i = 3.0 + 0.10X_i - 0.25Q_1 + 0.20Q_2 + 0.15Q_3$$

where quarter zero is the first quarter of 2009. Take the antilog of the appropriate coefficient from this equation and interpret the

- Y intercept, \hat{b}_0 .
- quarterly compound growth rate.
- second-quarter multiplier.

16.43 Refer to the exponential model given in Problem 16.42.

- What is the fitted value of the series in the fourth quarter of 2011?
- What is the fitted value of the series in the first quarter of 2011?
- What is the forecast in the fourth quarter of 2013?
- What is the forecast in the first quarter of 2014?

APPLYING THE CONCEPTS

 **16.44** The data in **Toys R Us** are quarterly revenues (in \$millions) for Toys R Us from 1996-Q1 through 2013-Q1. (Data extracted from *Standard & Poor's Stock Reports*, November 1995, November 1998, and April 2002, and Toys R Us, Inc., www.toysrus.com.)

- Do you think that the revenues for Toys R Us are subject to seasonal variation? Explain.
- Plot the data. Does this chart support your answer in (a)?
- Develop an exponential trend forecasting equation with quarterly components.
- Interpret the quarterly compound growth rate.

- Interpret the quarterly multipliers.

- What are the forecasts for 2013-Q2, 2013-Q3, 2013-Q4, and all four quarters of 2014?

16.45 Are gasoline prices higher during the height of the summer vacation season than at other times? The file **GasPrices** contains the mean monthly prices (in \$/gallon) for unleaded gasoline in the United States from January 2006 to May 2013. (Data extracted from U.S. Energy Information Administration, www.eia.gov/totalenergy/data/monthly/pdf/sec9_6.pdf.)

- Construct a time-series plot.
- Develop an exponential trend forecasting equation with monthly components.
- Interpret the monthly compound growth rate.
- Interpret the monthly multipliers.
- Write a short summary of your findings.

16.46 The data in **Travel** show the average traffic on Google recorded at the beginning of each month from January 2004 to August 2012 for searches from the United States concerning travel (scaled to the average traffic for the entire time period based on a fixed point at the beginning of the time period). (Data retrieved from Google Trends, www.google.com/trends, August 13, 2012.)

- Plot the time-series data.
- Develop an exponential trend forecasting equation with monthly components.
- What is the fitted value in August 2012?
- What are the forecasts for the last four months of 2012?
- Interpret the monthly compound growth rate.
- Interpret the July multiplier.

16.47 The file **CallCenter** contains the monthly call volume for an existing product. (Data extracted from S. Madadevan and J. Overstreet, "Use of Warranty and Reliability Data to Inform Call Center Staffing," *Quality Engineering* 24 (2012): 386–399.)

- Construct the time-series plot.
- Describe the monthly pattern in the data.
- In general, would you say that the overall call volume is increasing or decreasing? Explain.
- Develop an exponential trend forecasting equation with monthly components.
- Interpret the monthly compound growth rate.
- Interpret the January multiplier.
- What is the predicted call volume for month 60?
- What is the predicted call volume for month 61?
- How can this type of time-series forecasting benefit the call center?

16.48 The file **Silver-Q** contains the price in London for an ounce of silver (in US\$) at the end of each quarter from 2004 through 2012. (Data extracted from bit.ly/1afifi.)

- Plot the data.
- Develop an exponential trend forecasting equation with quarterly components.
- Interpret the quarterly compound growth rate.
- Interpret the first quarter multiplier.
- What is the fitted value for the last quarter of 2012?
- What are the forecasts for all four quarters of 2013?
- Are the forecasts in (f) accurate? Explain.

16.49 The file **Gold** contains the price in London for an ounce of gold (in US\$) at the end of each quarter from 2004 through 2012. (Data extracted from bit.ly/1afifi.)

- Plot the data.
- Develop an exponential trend forecasting equation with quarterly components.
- Interpret the quarterly compound growth rate.
- Interpret the first quarter multiplier.
- What is the fitted value for the last quarter of 2012?
- What are the forecasts for all four quarters of 2013?
- Are the forecasts in (f) accurate? Explain.

16.8 Index Numbers

An index number measures the value of an item (or group of items) at a particular point in time as a percentage of the value of an item (or group of items) at another point in time. The **Section 16.8 online topic** discusses this concept and illustrates its application.

THINK ABOUT THIS

When using a model, you must always review the assumptions built into the model and think about how novel or changing circumstances may render the model less useful.

Implicit in the time-series models developed in this chapter is that past data can be used to help predict the future. While using past data in this way is a legitimate application of time-series models, every so often, a crisis in financial markets illustrates that using models that rely on the past to predict the future is not without risk.

For example, during August 2007, many hedge funds suffered unprecedented losses.

Let the Model User Beware

Apparently, many hedge fund managers used models that based their investment strategy on trading patterns over long time periods. These models did not—and could not—reflect trading patterns contrary to historical patterns (G. Morgenson, “A Week When Risk Came Home to Roost,” *The New York Times*, August 12, 2007, pp. B1, B7). When fund managers in early August 2007 needed to sell stocks due to losses in their fixed income portfolios, stocks that were previously stronger became weaker, and weaker ones became stronger—the reverse of what the models expected. Making matters worse, many fund managers were using simi-

lar models and rigidly made investment decisions solely based on what those models said. These similar actions multiplied the effect of the selling pressure, an effect that the models had not considered and that therefore could not be seen in the models’ results.

This example illustrates that using models does not absolve you of the responsibility of being a thoughtful decision maker. Go ahead and use models—when appropriately used, they will enhance your decision making. But always remember that no model can completely remove the risk involved in making a business decision.

USING STATISTICS

Principled Forecasting, Revisited

In the Using Statistics scenario, you were the financial analyst for The Principled, a large financial services company. You needed to forecast movie attendance, revenues for Coca-Cola, and for Walmart to better evaluate investment opportunities for your clients.

For movie attendance, you used moving averages and exponential smoothing methods to develop forecasts. You predicted that the movie attendance in 2013 would be 1.37 billion.

For The Coca-Cola Company, you used least-squares linear, quadratic, and exponential models and autoregressive models to develop forecasts. You evaluated these alternative models and determined that the third-order autoregressive model gave the best forecast, according to several criteria. You predicted

that the revenue of The Coca-Cola Company would be \$47.2722 billion in 2013 and \$66.19 billion in 2014.

For Wal-Mart Stores, Inc., you used a least-squares regression model with seasonal components to develop forecasts. You predicted that Wal-Mart Stores would have revenues of \$112.3052 billion in the first quarter of fiscal 2014.

Given these forecasts, you now need to determine whether your clients should invest, and if so, how much they should invest in the movie industry or in The Coca-Cola Company or in Wal-Mart Stores, Inc.



SUMMARY

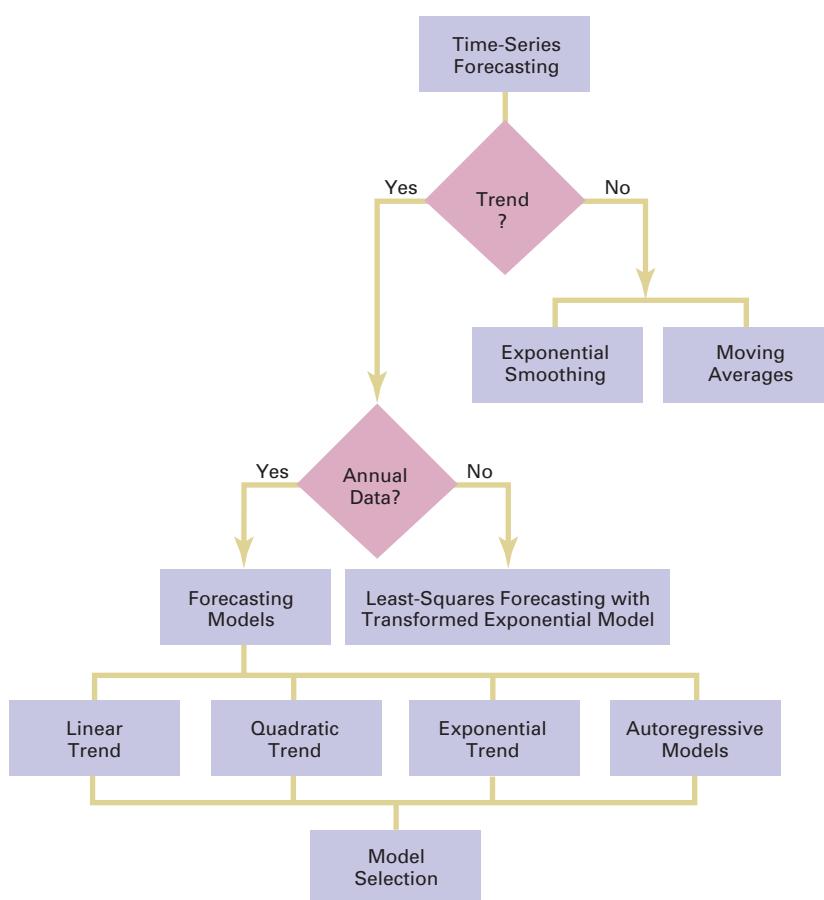
In this chapter, you studied smoothing techniques, least-squares trend fitting, autoregressive models, and forecasting of seasonal data. Figure 16.21 provides a summary chart for the time-series methods discussed in this chapter.

When using time-series forecasting, you need to plot the time series and answer the following question: Is there a trend in the data? If there is a trend, then you can use the autoregressive model or the linear, quadratic, or exponential

trend models. If there is no obvious trend in the time-series plot, then you should use moving averages or exponential smoothing to smooth out the effect of random effects and possible cyclical effects. After smoothing the data, if a trend is still not present, then you can use exponential smoothing to forecast short-term future values. If smoothing the data reveals a trend, then you can use the autoregressive model, or the linear, quadratic, or exponential trend models.

FIGURE 16.21

Summary chart of time-series forecasting methods



REFERENCES

1. Bowerman, B. L., R. T. O'Connell, and A. Koehler. *Forecasting, Time Series, and Regression*, 4th ed. Belmont, CA: Duxbury Press, 2005.
2. Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 1994.
3. Frees, E. W. *Data Analysis Using Regression Models: The Business Perspective*. Upper Saddle River, NJ: Prentice Hall, 1996.
4. Hanke, J. E., D. W. Wichern, and A. G. Reitsch. *Business Forecasting*, 7th ed. Upper Saddle River, NJ: Prentice Hall, 2001.
5. Microsoft Excel 2013. Redmond, WA: Microsoft Corp., 2012.
6. Minitab Release 16. State College, PA: Minitab Inc., 2010.

KEY EQUATIONS

Computing an Exponentially Smoothed Value in Time
Period i

$$E_1 = Y_1$$

$$E_i = WY_i + (1 - W)E_{i-1} \quad i = 2, 3, 4, \dots \quad (16.1)$$

Forecasting Time Period $i + 1$

$$\hat{Y}_{i+1} = E_i \quad (16.2)$$

Linear Trend Forecasting Equation

$$\hat{Y}_i = b_0 + b_1 X_i \quad (16.3)$$

Quadratic Trend Forecasting Equation

$$\hat{Y}_i = b_0 + b_1 X_i + b_2 X_i^2 \quad (16.4)$$

Exponential Trend Model

$$Y_i = \beta_0 \beta_1^{X_i} \varepsilon_i \quad (16.5)$$

Transformed Exponential Trend Model

$$\begin{aligned} \log(Y_i) &= \log(\beta_0 \beta_1^{X_i} \varepsilon_i) \\ &= \log(\beta_0) + \log(\beta_1^{X_i}) + \log(\varepsilon_i) \\ &= \log(\beta_0) + X_i \log(\beta_1) + \log(\varepsilon_i) \end{aligned} \quad (16.6)$$

Exponential Trend Forecasting Equation

$$\log(\hat{Y}_i) = b_0 + b_1 X_i \quad (16.7a)$$

$$\hat{Y}_i = \hat{\beta}_0 \hat{\beta}_1^{X_i} \quad (16.7b)$$

 p th-Order Autoregressive Models

$$Y_i = A_0 + A_1 Y_{i-1} + A_2 Y_{i-2} + \dots + A_p Y_{i-p} + \delta_i \quad (16.8)$$

First-Order Autoregressive Model

$$Y_i = A_0 + A_1 Y_{i-1} + \delta_i \quad (16.9)$$

Second-Order Autoregressive Model

$$Y_i = A_0 + A_1 Y_{i-1} + A_2 Y_{i-2} + \delta_i \quad (16.10)$$

 t Test for Significance of the Highest-Order Autoregressive Parameter, A_p

$$t_{STAT} = \frac{a_p - A_p}{S_{a_p}} \quad (16.11)$$

Fitted p th-Order Autoregressive Equation

$$\hat{Y}_i = a_0 + a_1 Y_{i-1} + a_2 Y_{i-2} + \dots + a_p Y_{i-p} \quad (16.12)$$

 p th-Order Autoregressive Forecasting Equation

$$\hat{Y}_{n+j} = a_0 + a_1 \hat{Y}_{n+j-1} + a_2 \hat{Y}_{n+j-2} + \dots + a_p \hat{Y}_{n+j-p} \quad (16.13)$$

Mean Absolute Deviation

$$MAD = \frac{\sum_{i=1}^n |Y_i - \hat{Y}_i|}{n} \quad (16.14)$$

Exponential Model with Quarterly Data

$$Y_i = \beta_0 \beta_1^{X_i} \beta_2^{Q_1} \beta_3^{Q_2} \beta_4^{Q_3} \varepsilon_i \quad (16.15)$$

Transformed Exponential Model with Quarterly Data

$$\begin{aligned} \log(Y_i) &= \log(\beta_0 \beta_1^{X_i} \beta_2^{Q_1} \beta_3^{Q_2} \beta_4^{Q_3} \varepsilon_i) \\ &= \log(\beta_0) + \log(\beta_1^{X_i}) + \log(\beta_2^{Q_1}) + \log(\beta_3^{Q_2}) \\ &\quad + \log(\beta_4^{Q_3}) + \log(\varepsilon_i) \\ &= \log(\beta_0) + X_i \log(\beta_1) + Q_1 \log(\beta_2) \\ &\quad + Q_2 \log(\beta_3) + Q_3 \log(\beta_4) + \log(\varepsilon_i) \end{aligned} \quad (16.16)$$

Exponential Growth with Quarterly Data Forecasting Equation

$$\log(\hat{Y}_i) = b_0 + b_1 X_i + b_2 Q_1 + b_3 Q_2 + b_4 Q_3 \quad (16.17)$$

Exponential Model with Monthly Data

$$Y_i = \beta_0 \beta_1^{X_i} \beta_2^{M_1} \beta_3^{M_2} \beta_4^{M_3} \beta_5^{M_4} \beta_6^{M_5} \beta_7^{M_6} \beta_8^{M_7} \beta_9^{M_8} \beta_{10}^{M_9} \beta_{11}^{M_{10}} \beta_{12}^{M_{11}} \varepsilon_i \quad (16.18)$$

Transformed Exponential Model with Monthly Data

$$\begin{aligned} \log(Y_i) &= \log(\beta_0 \beta_1^{X_i} \beta_2^{M_1} \beta_3^{M_2} \beta_4^{M_3} \beta_5^{M_4} \beta_6^{M_5} \beta_7^{M_6} \beta_8^{M_7} \beta_9^{M_8} \beta_{10}^{M_9} \beta_{11}^{M_{10}} \beta_{12}^{M_{11}} \varepsilon_i) \\ &= \log(\beta_0) + X_i \log(\beta_1) + M_1 \log(\beta_2) + M_2 \log(\beta_3) \\ &\quad + M_3 \log(\beta_4) + M_4 \log(\beta_5) + M_5 \log(\beta_6) + M_6 \log(\beta_7) \\ &\quad + M_7 \log(\beta_8) + M_8 \log(\beta_9) + M_9 \log(\beta_{10}) \\ &\quad + M_{10} \log(\beta_{11}) + M_{11} \log(\beta_{12}) + \log(\varepsilon_i) \end{aligned} \quad (16.19)$$

Exponential Growth with Monthly Data Forecasting Equation

$$\begin{aligned} \log(\hat{Y}_i) &= b_0 + b_1 X_i + b_2 M_1 + b_3 M_2 + b_4 M_3 + b_5 M_4 + b_6 M_5 \\ &\quad + b_7 M_6 + b_8 M_7 + b_9 M_8 + b_{10} M_9 + b_{11} M_{10} + b_{12} M_{11} \end{aligned} \quad (16.20)$$

KEY TERMS

autoregressive modeling 647
 causal forecasting methods 630
 cyclical effect 631
 exponential smoothing 634
 exponential trend model 640
 first-order autocorrelation 647
 first-order autoregressive model 648
 forecasting 630
 irregular effect 631

lagged predictor variable 647
 linear trend model 637
 mean absolute deviation (MAD) 656
 moving averages 632
 parsimony 656
 p th-order autocorrelation 647
 p th-order autoregressive model 647
 quadratic trend model 639

qualitative forecasting method 630
 quantitative forecasting method 630
 random effect 631
 seasonal effect 631
 second-order autocorrelation 647
 second-order autoregressive model 648
 time series 630
 time-series forecasting methods 630
 trend 631

CHECKING YOUR UNDERSTANDING

- 16.50** What is a time series?
- 16.51** What are the different components of a time-series model?
- 16.52** What is the difference between moving averages and exponential smoothing?
- 16.53** Under what circumstances is the exponential trend model most appropriate?
- 16.54** How does the least-squares linear trend forecasting model developed in this chapter differ from the least-squares linear regression model considered in Chapter 13?
- 16.55** How does autoregressive modeling differ from the other approaches to forecasting?
- 16.56** What are the different approaches to choosing an appropriate forecasting model?
- 16.57** What is the major difference between using S_{YX} and MAD for evaluating how well a particular model fits the data?
- 16.58** How does forecasting for monthly or quarterly data differ from forecasting for annual data?

CHAPTER REVIEW PROBLEMS

16.59 The data in the following table, stored in **Polio**, represent the annual incidence rates (per 100,000 persons) of reported acute poliomyelitis recorded over five-year periods from 1915 to 1955:

Year	1915	1920	1925	1930	1935	1940	1945	1950	1955
Rate	3.1	2.2	5.3	7.5	8.5	7.4	10.3	22.1	17.6

Source: Data extracted from B. Wattenberg, ed., *The Statistical History of the United States: From Colonial Times to the Present*, ser. B303.

- Plot the data.
- Compute the linear trend forecasting equation and plot the trend line.
- What are your forecasts for 1960, 1965, and 1970?
- Using a library or the Internet, find the actually reported incidence rates of acute poliomyelitis for 1960, 1965, and 1970. Record your results.
- Why are the forecasts you made in (c) not useful? Discuss.

16.60 The U.S. Department of Labor gathers and publishes statistics concerning the labor market. The file **Workforce** contains data on the size of the U.S. civilian noninstitutional population of people 16 years and over (in thousands) and the U.S. civilian noninstitutional workforce of people 16 years and over (in thousands) for 1984–2012. The workforce variable reports the number of people in the population who have a job or are actively looking for a job. (Data extracted from Bureau of Labor Statistics, U.S. Department of Labor, www.bls.gov.)

- Plot the time series for the U.S. civilian noninstitutional population of people 16 years and older.
- Compute the linear trend forecasting equation.
- Forecast the U.S. civilian noninstitutional population of people 16 years and older for 2013 and 2014.
- Repeat (a) through (c) for the U.S. civilian noninstitutional workforce of people 16 years and older.

16.61 The monthly wellhead and residential prices for natural gas (dollars per thousand cubic feet) in the United States from January 2008 through December 2012 are stored in **Natural Gas2**. (Data extracted from Energy Information Administration, U.S. Department of Energy, www.eia.gov, *Natural Gas Monthly*, May 31, 2013.)

For the wellhead price and the residential price,

- do you think the price for natural gas has a seasonal component?
- plot the time series. Does this chart support your answer in (a)?

- compute an exponential trend forecasting equation for monthly data.
- interpret the monthly compound growth rate.
- interpret the monthly multipliers. Do the multipliers support your answers in (a) and (b)?
- compare the results for the wellhead prices and the residential prices.

16.62 The data in the following table, stored in **McDonalds**, represent the gross revenues (in billions of current dollars) of McDonald's Corporation from 1975 through 2012:

Year	Revenues (\$billions)	Year	Revenues (\$billions)	Year	Revenues (\$billions)
1975	1.0	1988	5.6	2001	14.8
1976	1.2	1989	6.1	2002	15.2
1977	1.4	1990	6.8	2003	16.8
1978	1.7	1991	6.7	2004	18.6
1979	1.9	1992	7.1	2005	19.8
1980	2.2	1993	7.4	2006	20.9
1981	2.5	1994	8.3	2007	22.8
1982	2.8	1995	9.8	2008	23.5
1983	3.1	1996	10.7	2009	22.7
1984	3.4	1997	11.4	2010	24.1
1985	3.8	1998	12.4	2011	27.0
1986	4.2	1999	13.3	2012	27.6
1987	4.9	2000	14.2		

Source: Data extracted from *Moody's Handbook of Common Stocks*, 1980, 1989, and 1999; *Mergent's Handbook of Common Stocks*, Spring 2002; and "Investors: About McDonalds," www.mcdonalds.com.

- Plot the data.
- Compute the linear trend forecasting equation.
- Compute the quadratic trend forecasting equation.
- Compute the exponential trend forecasting equation.
- Determine the best-fitting autoregressive model, using $\alpha = 0.05$.
- Perform a residual analysis for each of the models in (b) through (e).
- Compute the standard error of the estimate (S_{YX}) and the MAD for each corresponding model in (f).
- On the basis of your results in (f) and (g), along with a consideration of the principle of parsimony, which model would you select for purposes of forecasting? Discuss.
- Using the selected model in (h), forecast gross revenues for 2013.

16.63 Teachers' Retirement System of the City of New York offers several types of investments for its members. Among the choices are investments with fixed and variable rates of return. There are several categories of variable-return investments. The Diversified Equity Fund consists of investments that are primarily made in stocks, and the Stable-Value Fund consists of investments in corporate bonds and other types of lower-risk instruments. The data in [TRSNYC](#) represent the value of a unit of each type of variable-return investment at the beginning of each year from 1984 to 2013. (Data extracted from "Historical Data-Unit Values, Teachers' Retirement System of the City of New York," [bit.ly/SESJF5](#).)

For each of the two time series,

- a. plot the data.
- b. compute the linear trend forecasting equation.
- c. compute the quadratic trend forecasting equation.
- d. compute the exponential trend forecasting equation.
- e. determine the best-fitting autoregressive model, using $\alpha = 0.05$.
- f. Perform a residual analysis for each of the models in (b) through (e).
- g. Compute the standard error of the estimate (S_{yx}) and the *MAD* for each corresponding model in (f).

h. On the basis of your results in (f) and (g), along with a consideration of the principle of parsimony, which model would you select for purposes of forecasting? Discuss.

- i. Using the selected model in (h), forecast the unit values for 2014.
- j. Based on the results of (a) through (i), what investment strategy would you recommend for a member of the Teachers' Retirement System of the City of New York? Explain.

REPORT WRITING EXERCISE

16.64 As a consultant to an investment company trading in various currencies, you have been assigned the task of studying long-term trends in the exchange rates of the Canadian dollar, the Japanese yen, and the English pound. Data from 1980 to 2012 are stored in [Currency](#), where the Canadian dollar, the Japanese yen, and the English pound are expressed in units per U.S. dollar.

Develop a forecasting model for the exchange rate of each of these three currencies and provide forecasts for 2013 and 2014 for each currency. Write an executive summary for a presentation to be given to the investment company. Append to this executive summary a discussion regarding possible limitations that may exist in these models.

CASES FOR CHAPTER 16

Managing Ashland MultiComm Services

As part of the continuing strategic initiative to increase subscribers to the *3-For-All* cable/phone/Internet services, the marketing department is closely monitoring the number of subscribers. To help do so, forecasts are to be developed for the number of subscribers in the future. To accomplish this task, the number of subscribers for the most recent 24-month period has been determined and is stored in [AMS16](#).

1. Analyze these data and develop a model to forecast the number of subscribers. Present your findings in a report

that includes the assumptions of the model and its limitations. Forecast the number of subscribers for the next four months.

2. Would you be willing to use the model developed to forecast the number of subscribers one year into the future? Explain.
3. Compare the trend in the number of subscribers to the number of new subscribers per month stored in [AMS13](#). What explanation can you provide for any differences?

Digital Case

Apply your knowledge about time-series forecasting in this Digital Case.

The *Ashland Herald* competes for readers in the Tri-Cities area with the newer *Oxford Glen Journal* (*OGJ*). Recently, the circulation staff at the *OGJ* claimed that their newspaper's circulation and subscription base is growing faster than that of the *Herald* and that local advertisers would do better if they transferred their advertisements from the *Herald* to the *OGJ*. The circulation department of the *Herald* has complained to the Ashland Chamber of Commerce about *OGJ*'s claims and has asked the chamber to investigate, a request that was welcomed by *OGJ*'s circulation staff.

Open [ACC_Mediation216.pdf](#) to review the circulation dispute information collected by the Ashland Chamber of Commerce. Then answer the following:

1. Which newspaper would you say has the right to claim the fastest-growing circulation and subscription base? Support your answer by performing and summarizing an appropriate statistical analysis.
2. What is the single most positive fact about the *Herald*'s circulation and subscription base? What is the single most positive fact about the *OGJ*'s circulation and subscription base? Explain your answers.
3. What additional data would be helpful in investigating the circulation claims made by the staffs of each newspaper?

CHAPTER 16 EXCEL GUIDE

EG16.1 The IMPORTANCE of BUSINESS FORECASTING

There are no Excel Guide instructions for this section.

EG16.2 COMPONENT FACTORS of TIME-SERIES MODELS

There are no Excel Guide instructions for this section.

EG16.3 SMOOTHING an ANNUAL TIME SERIES

Moving Averages

Key Technique Use the **AVERAGE**(*cell range that contains a sequence of L observed values*) function to compute moving averages and use the special worksheet value #N/A (not available) for time periods in which no moving average can be computed.

Example Compute the three- and five-year moving averages for the movie attendance data that is shown in Figure 16.3 on page 634.

In-Depth Excel Use the **COMPUTE worksheet** of the **Moving Averages workbook** as a template.

The worksheet already contains the data and formulas for the example. For other problems, paste the time-series data into columns A and B and adjust the moving average entries in columns C and D. (Open to the **COMPUTE_FORMULAS worksheet** to examine all formulas the worksheet uses.)

To construct a moving average plot for other problems, open to the adjusted COMPUTE worksheet and:

1. Select the cell range of the time-series data and the moving averages. (For the example, this cell range is **A1:D12**.)
2. Select **Insert → Scatter (X, Y) or Bubble Chart** (in older Excels **Scatter**) and select the **Scatter** gallery choice named **Scatter with Straight Lines and Markers**.
3. Relocate the chart to a chart sheet, turn off the gridlines, add axis titles, and modify the chart title by using the instructions in Appendix Section B.6.

Exponential Smoothing

Key Technique Use arithmetic formulas to compute exponentially smoothed values.

Example Compute the exponentially smoothed series ($W = 0.50$ and $W = 0.25$) for the movie attendance data that is shown in Figure 16.4 on page 635.

In-Depth Excel Use the **COMPUTE worksheet** of the **Exponential Smoothing workbook**, as a template.

The worksheet already contains the data and formulas for the example. In this worksheet, cells C2 and D2 contain the

formula $=B2$ that copies the initial value of the time series. The exponential smoothing begins in row 3, with cell C3 formula $=0.5 * B3 + 0.5 * C2$, and cell D3 formula $=0.25 * B3 + 0.75 * D2$. Note that in these formulas, the expression $1 - W$ in Equation (16.1) on page 634 has been simplified to the values 0.5 and 0.75, respectively. (Open to the **COMPUTE_FORMULAS worksheet** to examine all of the exponential smoothing formulas the worksheet uses.)

For other problems, paste the time-series data into columns A and B and adjust the exponentially smoothed entries in columns C and D. For problems with fewer than 11 time periods, delete the excess rows. For problems with more than 11 time periods, select row 12, right-click, and click **Insert** in the shortcut menu. Repeat as many times as there are new rows. Then select cell range **C11:D11** and copy the contents of this range down through the new table rows.

To construct a plot of exponentially smoothed values for other problems, open to the adjusted COMPUTE worksheet and:

1. Select the cell range of the time-series data and the exponentially smoothed values. (For the example, this cell range is **A1:D12**.)
2. Select **Insert → Scatter (X, Y) or Bubble Chart** (in older Excels **Scatter**) and select the **Scatter** gallery choice named **Scatter with Straight Lines and Markers**.
3. Relocate the chart to a chart sheet, turn off the gridlines, add axis titles, and modify the chart title by using the instructions in Appendix Section B.6.

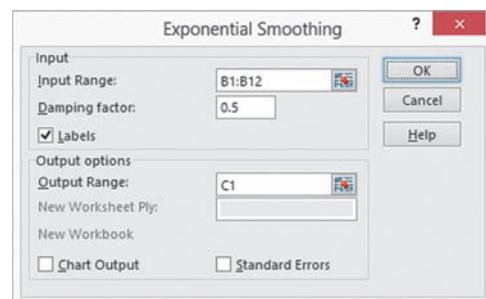
Analysis ToolPak

Use **Exponential Smoothing**. For the example, open to the **DATA worksheet** of the **Movie Attendance workbook** and:

1. Select **Data → Data Analysis**.
2. In the Data Analysis dialog box, select **Exponential Smoothing** from the **Analysis Tools** list and then click **OK**.

In the Exponential Smoothing dialog box (shown below):

1. Enter **B1:B12** as the **Input Range**.
2. Enter **0.5** as the **Damping factor**. (The damping factor is equal to $1 - W$.)
3. Check **Labels**, enter **C1** as the **Output Range**, and click **OK**.



In the new column C:

1. Copy the last formula in cell **C11** to cell **C12**.
2. Enter the column heading **ES(W = .50)** in cell **C1**, replacing the **#N/A** value.

To create the exponentially smoothed values that use a smoothing coefficient of $W = 0.25$, repeat steps 3 through 7 with these modifications: Enter **0.75** as the **Damping factor** in step 4, enter **D1** as the **Output Range** in step 5, and enter **ES(W = .25)** as the column heading in step 7.

EG16.4 LEAST-SQUARES TREND FITTING and FORECASTING

The Linear Trend Model

Modify the Section EG13.2 instructions (see page 638) to create a linear trend model. Use the cell range of the coded variable as the **X** variable cell range (called the **X Variable Cell Range** in the *PHStat* instructions, called the **cell range of X variable** in the *In-Depth Excel* instructions, and called the **Input X Range** in the *Analysis ToolPak* instructions). If you need to create coded values, enter them manually in a column. (If you have many coded values, you can use **Home** → **Fill** (in the Editing group) → **Series** and in the Series dialog box, click **Columns** and **Linear**, and select appropriate values for **Step value** and **Stop value**.)

The Quadratic Trend Model

Modify the Section EG15.1 instructions (see page 625) to create a quadratic trend model. Use the cell range of the coded variable and the squared coded variable as the **X** variables cell range (called the **X Variables Cell Range** in the *PHStat* instructions and the **Input X Range** in the *Analysis ToolPak* instructions). Use the Sections EG15.2 and EG15.1 instructions to create the squared coded variable and to plot the quadratic trend.

The Exponential Trend Model

Key Technique Use the **POWER(10, predicted log(Y))** function to compute the predicted **Y** values from the predicted **log(Y)** results.

To create an exponential trend model, first convert the values of the dependent variable **Y** to **log(Y)** values using the Section EG15.2 instructions on page 625. Then perform a simple linear regression analysis with residual analysis using the **log(Y)** values. Modify the Section EG13.5 “Residual Analysis” instructions using the cell range of the **log(Y)** values as the **Y** variable cell range and the cell range of the coded variable as the **X** variable cell range. (Note that the residual analysis instructions incorporate the Section EG13.2 “Determining the Simple Linear Regression Equation” instructions.)

Note the **Y** and **X** variable cell ranges are called the **Y Variable Cell Range** and **X Variable Cell Range** in the *PHStat* instructions, the **cell range of Y variable** and **cell range of X variable** in the *In-Depth Excel* instructions, and the **Input Y Range** and **Input X Range** in the *Analysis ToolPak* instructions.

If you use the *PHStat* or *In-Depth Excel* instructions, residuals will appear in a residuals worksheet. If you use the *Analysis ToolPak* instructions, residuals will appear in the **RESIDUAL**

OUTPUT area of the regression results worksheet. Because you use **log(Y)** values for the regression, the predicted **Y** and residuals listed are *log values* that need to be converted. [The *Analysis ToolPak* incorrectly labels the new column for the logs of the residuals as *Residuals*, and not as *LOG(Residuals)*, as you might expect.]

Convert the predicted **log(Y)** results to predicted **Y** results using the **POWER** function. Use an empty column in the residuals worksheet (*PHStat* or *In-Depth Excel*) or empty column ranges to the right of **RESIDUALS** OUTPUT area (*Analysis ToolPak*) to first add a column of formulas that use the **POWER** function to compute the predicted **Y** values. Then, add a second column that contains the original **Y** values. (Copy the original **Y** values to this column.) Finally, add a third new column that contains formulas in the form **= (revenue cell – predicted revenue cell)** to compute the actual residuals.

Use columns G through I of the **RESIDUALS worksheet** of the **Exponential Trend workbook** as a model. (Use the **Exponential Trend 2007 workbook** if you use an Excel version that is older than Excel 2010.) The worksheet already contains the values and formulas needed to create the Figure 16.10 plot that fits an exponential trend forecasting equation for The Coca-Cola Company revenues (see page 642).

To construct an exponential trend plot, first select the cell range of the time-series data and then use the Section EG2.5 instructions to construct a scatter plot. (For The Coca-Cola Company revenue example, use the cell range **B1:B18** in the **Data worksheet** of the **Coca-Cola workbook**.) Select the chart and

1. Select **Design** → **Add Chart Element** → **Trendline** → **More Trendline Options**.
2. In the Format Trendline pane, click **Exponential**.

If you use an Excel version older than Excel 2013, select **Layout** → **Trendline** → **More Trendline Options**. In the Format Trendline dialog box, click **Trendline Options** in the left pane and in the Trendline Options right pane, click **Exponential** and click **OK**.

Model Selection Using First, Second, and Percentage Differences

Use arithmetic formulas to compute the first, second, and percentage differences. Use division formulas to compute the percentage differences and use subtraction formulas to compute the first and second differences. Use the **COMPUTE worksheet** of the **Differences workbook**, shown in Figure 16.11 on page 645, as a model for developing a differences worksheet. (Open to the **COMPUTE_FORMULAS worksheet** to see all formulas used.)

EG16.5 AUTOREGRESSIVE MODELING for TREND FITTING and FORECASTING

Creating Lagged Predictor Variables

Create lagged predictor variables by creating a column of formulas that refer to a previous row’s (previous time period’s) **Y** value. Enter the special worksheet value **#N/A** (not available) for the cells in the column to which lagged values do not apply.

Use the **COMPUTE worksheet** of the **Lagged Predictors workbook**, similar to the Figure 16.13 Minitab worksheet on page 652 as a model for developing lagged predictor variables for

the first-order, second-order, and third-order autoregressive models. (Open to the **COMPUTE_FORMULAS** worksheet to see all formulas used.)

When specifying cell ranges for a lagged predictor variable, you include only rows that contain lagged values. Contrary to the usual practice in this book, you do not include rows that contain #N/A, nor do you include the row 1 column heading.

Autoregressive Modeling

Modify the Section EG14.1 instructions (see page 589) to create a third-order or second-order autoregressive model. Use the cell range of the first-order, second-order, and third-order lagged predictor variables as the *X* variables cell range for the third-order model. Use the cell range of the first-order and second-order lagged predictor variables as the *X* variables cell range for the second-order model (The *X* variables cell range is the **X Variables Cell Range** in the *PHStat* instructions and the **Input X Range** in the *Analysis ToolPak* instructions.) If using the *PHStat* instructions, modify step 3 to *clear not check First cells in both ranges contain label*. If using the *In-Depth Excel* instructions, use the **COMPUTE3** worksheet in lieu of the COMPUTE worksheet for the third-order model. If using the *Analysis ToolPak* instructions, do not check **Labels** in step 4.

Modify the Section EG13.2 instructions (see page 538) to create a first-order autoregressive model. Use the cell range of the first-order lagged predictor variable as the *X* variable cell range (called the **X Variable Cell Range** in the *PHStat* instructions, the *cell range of X variable* in the *In-Depth Excel* instructions, and the **Input X Range** in the *Analysis ToolPak* instructions). If using the *PHStat* instructions, modify step 3 to *clear not check First cells in both ranges contain label*. If using the *Analysis ToolPak* instructions, do not check **Labels** in step 4.

EG16.6 CHOOSING an APPROPRIATE FORECASTING MODEL

Performing a Residual Analysis

To create residual plots for the linear trend model or the first-order autoregressive model, use the instructions in Section EG13.5 on page 539. To create residual plots for the quadratic trend model or second-order autoregressive model, use the instructions in Section EG14.3 on page 590. To create residual plots for the exponential trend model, use the instructions in Section EG16.4 on page 670. To create residual plots for the third-order autoregressive model, use the instructions in Section EG14.3 on page 590 but use the **RESIDUALS3** worksheet instead of the RESIDUALS worksheet if using the *In-Depth Excel* instructions.

Measuring the Magnitude of the Residuals Through Squared or Absolute Differences

To compute the mean absolute deviation (*MAD*), first perform a residual analysis. Then add a formula in the form =**SUMPRODUCT(ABS(cell range of residual values)) / COUNT(cell range of the residual values)**. In the *cell range of the residual values* do not include the column heading as is the

standard practice in this book. (See Appendix Section F.4 to learn more about the application of **SUMPRODUCT** function in this formula.)

The **RESIDUALS_FORMULAS** worksheet of the **Exponential Trend** workbook shows an example of this formula in cell I19 for The Coca-Cola Company revenues example.

A Comparison of Four Forecasting Methods

Construct a model comparison worksheet similar to the one shown in Figure 16.18 on page 657 by using **Paste Special values** (see Appendix Section B.4) to transfer results from regression results worksheets. For the *SSE* values (row 20 in Figure 16.8), copy the regression results worksheet cell C13, the *SS* value for Residual in the ANOVA table. For the *S_{YX}* values (row), copy the regression results worksheet cell B7, labeled Standard Error, for all but the exponential trend model. For the *MAD* values, add formulas as discussed in the previous section.

For the *S_{YX}* value for the exponential trend model, enter a formula in the form =**SQRT(exponential SSE cell / (COUNT(cell range of exponential residuals) - 2))**. For the Figure 16.18 worksheet, this formula is =**SQRT(H20 / (COUNT(H3:H19) - 2))**. Use the **COMPARE** worksheet of the **Forecasting Comparison** workbook as a model. Open to the **COMPARE_FORMULAS** worksheet to examine all formulas. This worksheet also shows an alternative way that uses a formula to display the *SSE* values.

EG16.7 TIME-SERIES FORECASTING of SEASONAL DATA

Least-Squares Forecasting with Monthly or Quarterly Data

To develop a least-squares regression model for monthly or quarterly data, add columns of formulas that use the **IF** function (see Appendix Section F.4) to create dummy variables for the quarterly or monthly data. Enter all formulas in the form =**IF(comparison, 1, 0)**.

Shown below are the first five rows of columns F through K of a data worksheet that contains dummy variables. In the first illustration, columns F, G, and H contain the quarterly dummy variables Q1, Q2, and Q3 that are based on column B coded quarter values (not shown). In the second illustration, columns J and K contain the two monthly variables M1 and M6 that are based on column C month values (also not shown).

	F	G	H
1	Q1	Q2	Q3
2	=IF(B2 = 1, 1, 0)	=IF(B2 = 2, 1, 0)	=IF(B2 = 3, 1, 0)
3	=IF(B3 = 1, 1, 0)	=IF(B3 = 2, 1, 0)	=IF(B3 = 3, 1, 0)
4	=IF(B4 = 1, 1, 0)	=IF(B4 = 2, 1, 0)	=IF(B4 = 3, 1, 0)
5	=IF(B5 = 1, 1, 0)	=IF(B5 = 2, 1, 0)	=IF(B5 = 3, 1, 0)

	J	K
1	M1	M6
2	=IF(C2 = "January", 1, 0)	=IF(C2 = "June", 1, 0)
3	=IF(C3 = "January", 1, 0)	=IF(C3 = "June", 1, 0)
4	=IF(C4 = "January", 1, 0)	=IF(C4 = "June", 1, 0)
5	=IF(C5 = "January", 1, 0)	=IF(C5 = "June", 1, 0)

CHAPTER 16 MINITAB GUIDE

MG16.1 THE IMPORTANCE of BUSINESS FORECASTING

There are no Minitab Guide instructions for this section.

MG16.2 COMPONENT FACTORS of TIME-SERIES MODELS

There are no Minitab Guide instructions for this section.

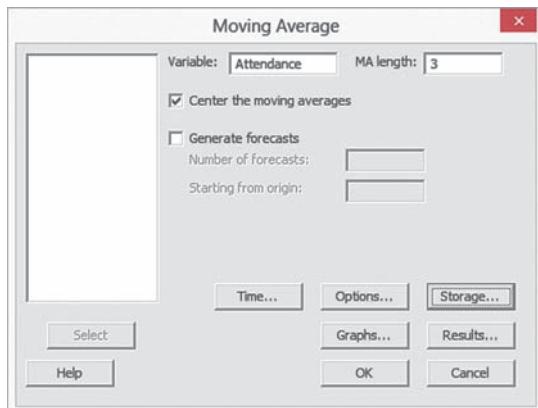
MG16.3 SMOOTHING an ANNUAL TIME SERIES

Moving Averages

Use Moving Average.

For example, to compute the Figure 16.3 moving averages shown on page 634, open to the **Movie Attendance worksheet**. Select **Stat → Time Series → Moving Average**. In the Moving Average dialog box (shown below):

1. Double-click **C2 Attendance** in the variables list to add **Attendance** to the **Variable** box.
2. Enter **3** in the **MA length** box.
3. Check **Center the moving averages**.
4. Click **Storage**.



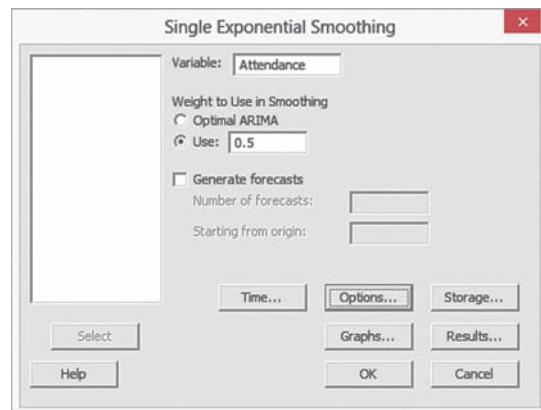
5. In the Moving Average – Storage dialog box (not shown), check **Moving Averages** and then click **OK**.
6. Back in the Moving Average dialog box, click **Graphs**.
7. In the Moving Average – Graphs dialog box (not shown), click **Plot smoothed vs. actual** and clear all check boxes, and then click **OK**.
8. Back in the Moving Average dialog box, click **Results**.
9. In the Moving Average - Results dialog box (not shown), click **Summary table and results table** and then click **OK**.
10. Back in the Moving Average dialog box, click **OK**.
11. Enter **MA 3-Yr** as the name for **column C3** (replacing **AVER1**).

To add the five-year moving averages, repeat steps 1 through 10, entering **5** in the **MA length** box in step 2. Then enter **MA 5-Yr** as the name for **column C4** (replacing **AVER1**).

Exponential Smoothing

Use **Single Exp Smoothing** to compute exponential smoothed values. For example, to compute the Figure 16.4 exponential smoothed values shown on page 635, open to the **Movie Attendance worksheet**. Select **Stat → Time Series → Single Exp Smoothing**. In the Single Exponential Smoothing dialog box (shown below):

1. Double-click **C2 Attendance** in the variables list to add **Attendance** to the **Variable** box.
2. Click **Use** and enter **0.50** in its box (for a *W* value of 0.50).
3. Click **Options**.



4. In the Single Exponential Smoothing - Options dialog box, enter **1** in the **Use average of first K observations** box and then click **OK**.
5. Back in the Moving Average dialog box, click **Storage**.
6. In the Single Exponential Smoothing – Storage dialog box (not shown), check **Smoothed data** and then click **OK**.
7. Back in the Moving Average dialog box, click **Graphs**.
8. In the Moving Average – Graphs dialog box (not shown), click **Plot smoothed vs. actual** and clear all check boxes, and then click **OK**.
9. Back in the Single Exponential Smoothing dialog box, click **Results**.
10. In the Single Exponential Smoothing - Results dialog box, click **Summary table and results table** and then click **OK**.
11. Back in the Single Exponential Smoothing dialog box, click **OK**.
12. Enter **ES(W = 0.50)** as the name for **column C3** (replacing **SMOO1**).

For a W value of 0.25, repeat steps 1 through 11, entering **0.25** in step 2. Then enter **ES(W = 0.25)** as the name for **column C4** (replacing SMOO1).

MG16.4 LEAST-SQUARES TREND FITTING and FORECASTING

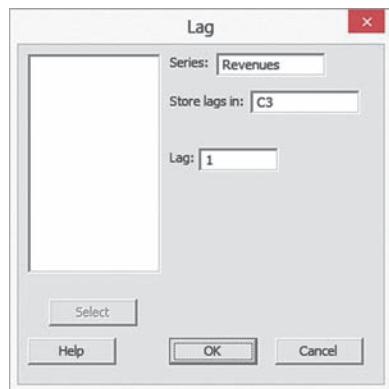
In Chapters 13 through 15, you used Minitab for the simple linear regression model and for a variety of multiple regression models. In this chapter, time-series models were developed assuming either a linear, quadratic, or exponential trend. For the linear trend model, see Section MG13.2 on page 541. For the quadratic and the exponential trend models, see Sections MG15.1 and MG15.2 on pages 626–627.

MG16.5 AUTOREGRESSIVE MODELING for TREND FITTING and FORECASTING

Creating Lagged Predictor Variables

Use **Lag** to create lagged predictor variables for autoregressive models. For example, to create the Figure 16.13 lagged variables worksheet on page 634, open to the **Coca-Cola worksheet**. Select **Stat → Time Series → Lag**. In the Lag dialog box (shown below):

1. Double-click **C3 Revenues** in the variables list to add **Revenues** to the **Series** box.
2. Enter **C3** in the **Store lags in** box and press **Tab**.
3. Enter **1** in the **Lag** box (for a one-period lag).
4. Click **OK**.



5. In the worksheet, enter **Lag1** as the name for **column C3**.
6. Again select **Stat → Time Series → Lag**. In the Lag dialog box, enter **C4** in the **Store lags in** box, press **Tab**, and enter **2** in the **Lag** box (for a 2-period lag). Click **OK**.
7. In the worksheet, enter **Lag2** as the name for **column C4**.
8. Reselect **Stat → Time Series → Lag**. In the Lag dialog box, enter **C5** in the **Store lags in** box, press **Tab**, and enter **3** in the **Lag** box (for a 3-period lag). Click **OK**.
9. In the worksheet, enter **Lag3** as the name for **column C5**.

Autoregressive Modeling

Modify the Section MG14.1 “Interpreting the Regression Coefficients” instructions (see page 592) to create a third-order or second-order autoregressive model. Add the names of the columns containing the first-order, second-order, and third-order lagged predictor variables to the **Predictors** box for the third-order model. Add the names of the columns containing the first-order, and second-order lagged predictor variables to the **Predictors** box for the second-order model.

Modify the Section MG13.2 instructions (see page 541) to create a first-order autoregressive model. In step 2, add the name of the column containing the first-order lagged predictor variable to the **Predictors** box.

MG16.6 CHOOSING an APPROPRIATE FORECASTING MODEL

A Comparison of Four Forecasting Methods

When you compare the four forecasting models, you use residual analysis to examine the models. Use the instructions in Section MG13.5 on page 541 to create residual plots for the linear trend model or first-order autoregressive models. Use the instructions in Section MG14.1 on page 592 to create residual plots for the quadratic and the exponential trend models.

MG16.7 TIME-SERIES FORECASTING of SEASONAL DATA

Least-Squares Forecasting with Monthly or Quarterly Data

Use **Calculator** to create dummy variables for the quarterly or monthly data. For example, to create the dummy quarterly variable Q1 for the Table 16.3 Wal-Mart Stores quarterly revenues on page 658, open to the **WalMart worksheet**. Select **Calc → Calculator**. In the Calculator dialog box:

1. Enter **C5** in the **Store result in variable** box.
2. Enter **IF(Quarter = 1,1,0)** in the **Expression** box.
3. Click **OK**.
4. Enter **Q1** as the name for **column C5**.

In step 2, use the expression **IF(Quarter = 2,1,0)** to create the dummy quarterly variable Q2 or use **IF(Quarter = 3,1,0)** to create the quarterly variable Q3.

For monthly variables, first convert the values to text, if necessary, using **Change Data Type** (see Section MG1.1 on page 34). Then select **Calc → Calculator** and enter expressions such as **IF(Month = "January",1,0)** in the **Expression** box.

CHAPTER

17

Business Analytics

CONTENTS

- 17.1 Descriptive Analytics
- 17.2 Predictive Analytics
- 17.3 Classification and Regression Trees
- 17.4 Neural Networks
- 17.5 Cluster Analysis
- 17.6 Multidimensional Scaling

USING STATISTICS: Finding the Right Lines at WaldoLands, Revisited

CHAPTER 17 SOFTWARE GUIDE

OBJECTIVES

- To develop dashboard elements such as sparklines, gauges, bullet graphs, and treemaps for descriptive analytics.
- To learn how to use classification and regression trees for predictive analytics
- To learn how to use neural nets for predictive analytics
- To learn how to use cluster analysis for predictive analytics
- To learn how to use multidimensional scaling for predictive analytics

USING STATISTICS

Finding the Right Lines at WaldoLands*

The managers of WaldoLands, the theme park that licenses the characters from the Waldowood stories, seek to stabilize and grow their business. Last tourist season, their park was plagued by a number of major ride breakdowns, long lines at popular attractions and key food service areas, and a general inability to respond to the park's day-to-day operating status.

Last year's problems led to numerous unfavorable reviews in key social media travel websites and the managers are concerned that possible patrons may decide to visit competing parks run by Universal Parks & Resorts and Six Flags Entertainment. For this year, the managers have added the LineJumper service that allows patrons to "jump" to the head of a line and are offering the premium-priced No-Stress-Express experience that offers special guided tours and behind-the-scenes access. The managers also hope the new multimillion-dollar Rabbit Creek Racers and a greatly expanded *MirrorGate* Experience, based on a popular sci-fi franchise, will boost attendance, even as the managers fret about the technical complexity of these rides.

As part of the general management team, you puzzle over how to manage the park to maximize both revenues and customer satisfaction. You have also been asked to ensure that guests at the adjacent Waldowood Resort & Convention Center have opportunities for an enhanced theme park experience. How can you achieve these goals while managing an enterprise that is the size of a small city?



Arinahabich/Fotolia

*The contents, descriptions, and characters of WaldoLands and Waldowood are copyright © 2014, 2011 Waldowood Productions, and used with permission.

For many chapters now, you have been reading about methods that make inferences about population data. Many of these techniques were developed 100 or more years ago and were first made practical by the use of sample data. Other techniques awaited computerization to make them practical and widely used.

Today, business statistics continues to evolve. In this book, you have already read how to use visualizations as a means of preliminary analysis. Such a technique was not practical until recently when scatter plots and the like could be constructed in the “blink of an eye” on cheap, everyday display devices. (Such plots once required expensive, large, special-purpose plotters and took minutes to construct.) As business people gain the ability to retrieve and process larger amounts of data in smaller amounts of time, sometimes approaching *near real time*,¹ some have asked at what point does the need for using *samples* to expedite analysis disappear? Might there not be a day when business decision makers could just analyze all of the data continuously as it flows into the business in near real time?

While that day of continuous data analysis has not yet arrived in most cases, these questions taken together have created the demand for methods known collectively as **business analytics**. Analytics represents an evolution of preexisting statistical methods combined with advances in information systems and techniques from management science. Analytics is naturally interdisciplinary and this nature underscores how statistics is an important part of your business education, one of the themes of the Getting Started chapter.

Descriptive analytics, predictive analytics, and prescriptive analytics form the three broad categories of analytic methods. **Descriptive analytics** explores business activities that have occurred or are occurring in the present moment. **Predictive analytics** identifies what is likely to occur in the (near) future and finds relationships in data that may not be readily apparent using descriptive analytics. **Prescriptive analytics** investigates what should occur and prescribes the best course of action for the future. Predictive and prescriptive analytics make practical the use of *big data* (see page 4) to support decision making, although many of these techniques also work with smaller sets of data, as examples in this chapter demonstrate. This chapter begins with *descriptive analytics* but focuses on *predictive analytics*. The chapter does not cover prescriptive analytics methods.

17.1 Descriptive Analytics

In Chapters 2 and 3, you learned descriptive methods that organized and visualized data that had been previously collected. What if you could combine, collect, organize, and visualize tasks of the DCOVA framework into a single task that could be done in near real-time? That would change the historical nature of descriptive methods—your summaries represent the status of a business activity *in the past*—into a tool that could be used for day-to-day, if not minute-by-minute, business monitoring in the *present*. Giving decision makers this ability is one of the goals of descriptive analytics.

Being able to do real-time monitoring can be useful for a business that handles a *perishable inventory*. Perishable inventory is inventory that will disappear after a particular event takes place such as an airplane taking off for its destination or the end of a concert. Empty seats on the airplane or at the concert cannot be sold at a later time. Perishable inventory also occurs with less tangible inventory such as spaces reserved for advertisements on a commercial web page—such spaces cannot be sold after the page has been viewed. In the past, the problem of perishable inventory was handled by models that predicted consumer behavior based on historical patterns. A concert promoter set prices based on the best *estimation* of ticket-buying behavior. Today, by constantly monitoring sales, the promoter can use a *dynamic pricing model* in which the price of tickets could fluctuate in near real time based on whether sales are exceeding or failing to meet predicted demand.

Real-time monitoring can also be useful for a business that manages *flows* of people or objects that can be adjusted in near real time, especially when there are more than one flow and the flows are interrelated. For example, overseers of a large sports stadium could benefit from monitoring the flows of cars in parking facilities as well as the flow of fans into the stadium and

¹Near real time in this sense means in a time period short enough that the results produced can immediately affect operational management decision making.

While many assembly-line processes such as the cereal-filling line in the Oxford Cereals case (see Chapter 7) are flows, such business processes may not be good candidates for real-time adjustments for reasons explained in the online Chapter 19.

redirect stadium personnel to assist at points of congestion. In the WaldoLands scenario, managers could monitor flows of patrons through the ticket booths and into the theme park while also keeping an eye on the length of waiting lines and the use of the LineJumper service. This would allow the managers to adjust ride lengths or dispatch live performers to entertain patrons in line and to try to redirect patrons to areas of the park that are currently under capacity.

Dashboards

Over several decades, people talked about developing *executive information systems* that would put information at the “fingertips” of decision makers. Many of these efforts have spurred the development of dashboards that use descriptive analytics methods to present up-to-the-minute operational status about a business.

An analytic **dashboard** provides this information in a visual form that is intended to be easy to comprehend and review. Dashboards can contain the summary tables and charts discussed in Chapter 2, as well as newer or more novel forms of information presentation, that can summarize big data as well as smaller sets of data. The dashboard in Figure 17.1 displays key WaldoLands operational statistics that are updated on a near-real-time basis. Clicking one of the categories would lead to other displays that contain additional information about theme park operations.

FIGURE 17.1
A WaldoLands dashboard

LEARN MORE

Learn how this dashboard could be enhanced in the SHORT TAKES for Chapter 17.



Sparklines are one of the descriptive analytic methods that dashboards can contain. **Sparklines** summarize time-series data as small, compact graphs designed to appear as part of a table (or a written passage). In Figure 17.2, sparklines display the wait times for WaldoLands attractions at half-hour intervals for the current day, helping to provide context for the current wait times that are indicated by the dot markers. For example, the sparkline for the Rabbit Springs Racers ride shows that the current wait time is one of the longest wait times for the day.

FIGURE 17.2
WaldoLands wait times table with sparklines

Student Tip

To be used properly, a set of sparklines must share the same X and Y axes, a requirement that some programs, including Microsoft Excel, do not enforce.

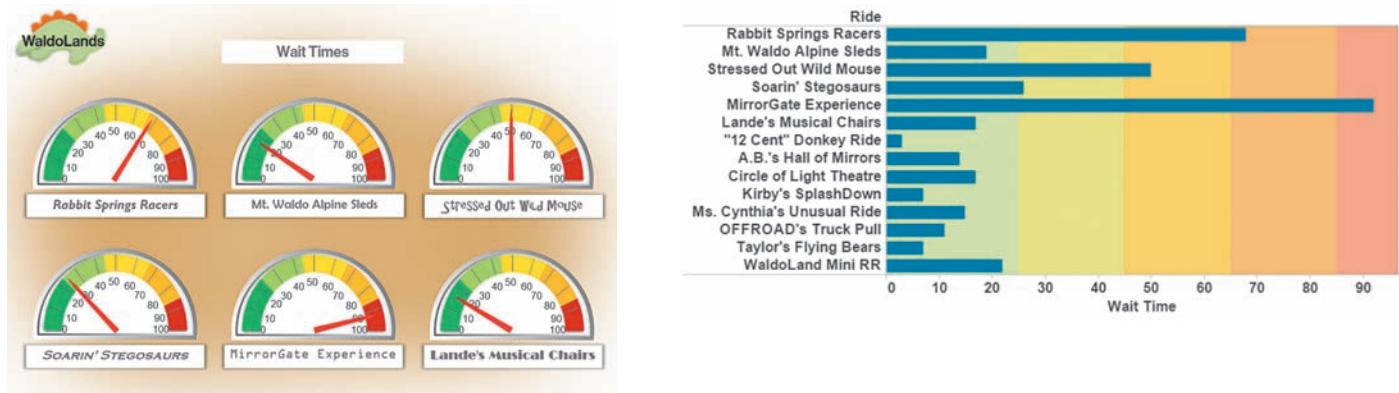
	A	B	C
1	WaldoLands Wait Times		
2			
3	Ride	Wait Times	
		Today	Current
4	Rabbit Springs Racers		82
5	Mt. Waldo Alpine Sleds		28
6	Stressed Out Wild Mouse		25
7	Soarin' Stegosaurs		63
8	MirrorGate Experience		42
9	Lande's Musical Chairs		23
10	OFFROAD's Truck Pull		9
11	Circle of Light Theatre		12
12	Kirby's SplashDown		26
13	Taylor's Flying Bears		2
14	Ms. Cynthia's Unusual Ride		11
15	"12 Cent" Donkey Ride		9
16	A.B.'s Hall of Mirrors		18
17	WaldoLand Mini RR		30

²This tension between what decision makers might find visually appealing and what statisticians and information specialists have found most useful reflects the relative newness of these descriptive methods. Over time, this tension may ease and an acceptable standard for representing such information may emerge.

Analogous to automotive dashboards, analytic dashboards can provide warnings when predefined conditions are met or exceeded. Figure 17.3 contains a set of **gauges** and a **bullet graph** that both display the wait-line status for WaldoLands attractions. These displays combine a single numerical measure (wait time) with one of five categorical values that rates the wait time subjectively, from excellent (less than 25 minutes) to poor (more than 85 minutes). While gauges have been a popular choice in business, most information design specialists prefer bullet graphs because those graphs foster the direct comparison of each measurement (wait time in Figure 17.3). Gauges can also consume a lot of visual space in a dashboard. For example, in Figure 17.3, note the amount of the space the gauges consume to show the status of the six most popular rides. The corresponding bullet graph can display the status of 14 rides and present the wait times in a way that facilitates comparisons. For these reasons, some consider gauges little more than examples of chartjunk (see reference 4), even as many decision makers request them due to their visual appeal.²

FIGURE 17.3

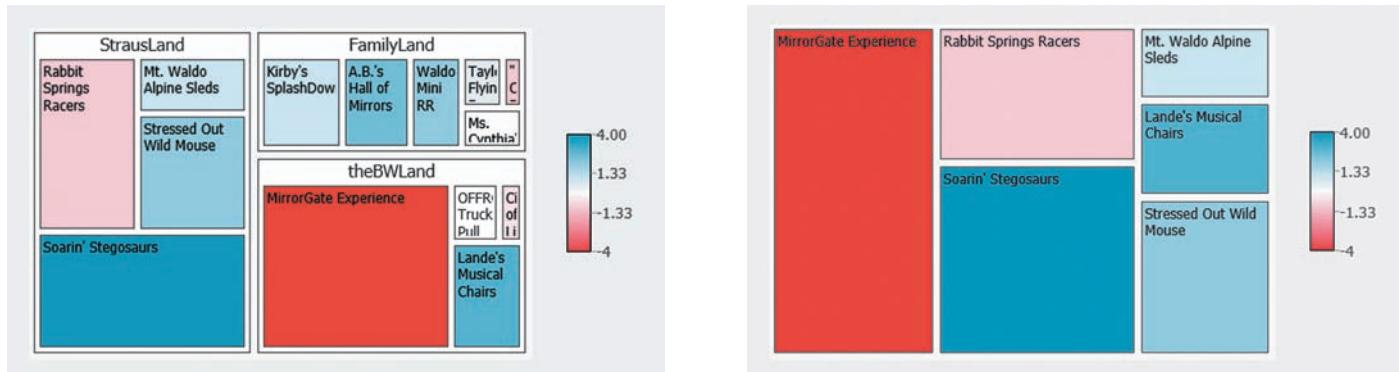
Gauges and bullet graph of wait times for WaldoLands attractions



Dashboards may also contain **treemaps** that help visualize two variables, one of which must be categorical. Treemaps are especially useful when categories can be grouped to form a multilevel hierarchy or *tree*. Figure 17.4 displays a pair of treemaps that visualize the number of social media comments made today about WaldoLands attractions (the size of each rectangle) and gives an assessment of the average comment's favorability (the color), from very positive (dark blue) to very negative (dark red). The left treemap shows each ride grouped by the “land” of WaldoLands (StrausLand, the BWLand, or FamilyLand) where the attraction is found. The right treemap shows the data for the six most popular WaldoLands attractions, illustrating that treemaps can be used with nonhierarchical information as well.

FIGURE 17.4

Treemaps of number and favorability of social media comments about WaldoLands attractions



Student Tip

Treemaps are especially effective at displaying nested geographical data, such as market preferences or voting patterns by state and county, as they eliminate the distortion that placing such data on political (geographic) maps can create because of differences in the sizes of the political units. For example, a treemap would eliminate the distortion that always makes sparsely populated Montana much more prominent than the densely populated New Jersey on a U.S. political map.

When combined with the Figure 17.3 gauges or bullet graph, the treemap at right in Figure 17.4 would allow managers to preliminarily conclude that the negativity of comments seems to be tied to current wait lines and that rides with the shortest wait lines may generate the fewest social media comments. These relationships could then be further investigated and, if the former one was confirmed, managers could, in the future, respond to excessive wait lines by shortening the ride length to handle more customers, by sending live performers to entertain those waiting in line, or by instructing park staff to divert incoming park patrons to other rides.

Note that gauges, bullet graphs, and treemaps use color to represent the value of a second variable, thereby increasing the data density of the displays, one of the principles of good information design (see reference 11). However, when using these displays, particularly bullet graphs, and treemaps, avoid using color spectrums that run from red to green, the two colors most subject to confusion due to color vision deficiencies. (This is less of a problem with gauges, as colors subject to confusion will have unique positions on the gauge dial.)

There is a tradeoff using treemaps when the second variable is *numerical* and has been recoded as a color or shade on a color spectrum (as was done in Figure 17.4). You must always consider whether the positive effects of making a preliminary analysis through visual means offsets the loss of the preciseness of the numerical values. While this tradeoff is always favorable when summarizing sets of “big data,” the advantage narrows as the set of data shrinks. (Thus could be used as an argument against constructing the treemap of the six most popular rides shown in Figure 17.4.)

Data Discovery

Data discovery methods allow decision makers to interactively organize or visualize data and perform preliminary analyses. These methods can be used to take a closer look at historical or status data, to quickly review data for unusual values or outliers, or to construct visualizations for management presentations. In these ways, data discovery realizes the earlier promise of executive information systems to give decision makers the tools of data exploration and presentation.

In its simplest version, data discovery involves **drill-down**, the revealing of the data that underlies a higher-level summary. For example, clicking the merchandise entry in the Figure 17.1 WaldoLands dashboard would reveal more detailed information such as the table of sales by “lands” shown in the left table in Figure 17.5. In turn, this summary can be drilled down to reveal sales by each store in the theme park (see right table in Figure 17.5). At this level of detail, sales at Peri’s Playtime are significantly lower than the other stores, perhaps suggesting that this store be closed, relocated, or have its merchandise mix reconsidered.

FIGURE 17.5

WaldoLands merchandise sales summarized on two different levels

Land	Merchandise Sales
FamilyLand	25613
StrausLand	17432
theBWLand	16103
Grand Total	59148

B	C
Land	Merchandise Sales
FamilyLand	25613
Egbert's Cards & Gifts	3655
Ms. Cynthia's Store	2957
Peri's Playtime	1497
Taylor's T-Shirts	7847
Tomoko's Closet	4063
Waldo's Green Things	5594
StrausLand	17432
All Things Mice!	2985
Super Dino Bros Games & Tricks	5138
Ties & Vests by Straus	2782
Waldo's Towels & Blankets	6527
theBWLand	16103
Dirk "Sonny" Lande's Music Emporium	3243
theBW Official Store	8414
Trevor & Tyler's Treasures	4446
Grand Total	59148

Another level of drill-down (not shown) would reveal the sales of each item or SKU (stock-keeping unit) sold in each store. By reorganizing that list by item, WaldoLands managers could discover which items are selling the best and may be subject to stockouts.

Drill-down can also reveal the data for variables not initially displayed in a summary table or chart. For example, Figure 2.16 on page 69 displays a multidimensional contingency table for the sample of 316 retirement funds used in earlier chapters. This Excel PivotTable also demonstrates one level of drill-down, drilling down the type of fund by market cap. Further drill-down is possible by clicking one of the cells. For example, clicking the cell that contains 2.85%, the percentage of funds that are small market cap value funds with low risk, creates the Figure 17.6 worksheet that contains the details for the 9 retirement funds that the 2.85% represents.

FIGURE 17.6

Results of drilling down to the details about small market cap value funds with low risk

A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Fund Number	Market Cap	Type	Assets	Turnover Ratio	Beta	SD	Risk	1YrReturn%	3YrReturn%	5YrReturn%	10YrReturn%	Expense Ratio	Star Rating
2	RF316	Small	Value	71.30	14.00	0.84	13.79	Low	4.83	7.12	4.41	9.80	1.27	Four
3	RF310	Small	Value	664.50	68.00	0.71	11.68	Low	8.87	9.63	11.35	11.51	1.46	Five
4	RF308	Small	Value	48.30	14.60	0.94	17.02	Low	11.79	10.40	-2.27	4.27	1.66	Two
5	RF306	Small	Value	40.90	28.00	1.16	18.97	Low	12.49	11.08	5.11	8.76	1.60	Three
6	RF304	Small	Value	73.30	32.00	1.15	18.69	Low	22.54	11.76	6.27	10.15	1.61	Three
7	RF303	Small	Value	103.20	16.78	1.05	17.41	Low	16.54	12.09	7.31	8.29	1.51	Four
8	RF302	Small	Value	1837.60	16.04	1.20	1.92	Low	13.78	12.11	4.91	10.10	1.38	Four
9	RF300	Small	Value	1980.30	27.00	1.14	18.80	Low	20.13	13.13	6.63	9.63	1.13	Four
10	RF298	Small	Value	127.80	89.00	0.95	15.90	Low	7.35	14.69	3.09	9.86	1.50	Four

Data discovery also allows decision makers to add or remove variables or statistics to uncover new patterns in the data. Replacing the count of funds with the mean of the ten-year return percentage to the retirement funds PivotTable reveals that many of the highest percentage gains over the ten-year period come from small market cap funds, particularly those funds with low risk and that small market cap value funds with low risk had the highest mean ten-year percentage return (see Figure 17.7). In addition, mid-cap funds had a higher return than large cap funds for both growth and value funds and also for low and average risk funds.

FIGURE 17.7

Mean 10-year return percentages for the sample of 316 retirement funds, by fund type, market cap, and risk

Type	Low	Average	High	Grand Total
Growth	7.13	8.17	6.65	7.45
Large	6.37	7.16	8.42	6.51
Mid-Cap	8.36	8.65	0.00	8.50
Small	8.59	7.94	5.89	7.86
Value	6.69	8.11	6.22	6.95
Large	5.81	7.01	4.03	5.87
Mid-Cap	7.87	8.23	0.00	7.94
Small	9.15	8.55	7.32	8.70
Grand Total	6.99	8.16	6.55	7.31

Some data discovery methods are primarily visual. Perhaps the simplest method is the direct manipulation of a visual, as was done with the three-dimensional scatter plot shown in Figure 14.1 on page 545. With map-type data, zooming, the visual equivalent of drill-down is also possible. For example, the websites such as **GasBuddy.com** and **GasPriceWatch.com** display fuel prices at gas stations as a series of maps that can be zoomed down to see specific locations.

Data discovery can also provide a means of preliminary analysis. For example, Figure 17.8 displays several Excel **slicers**, panels of clickable buttons that appear superimposed over a worksheet. Each slicer panel corresponds to one of the variables that is under study, and each button in a variable's slicer panel represents a unique value of the variable that is found in the data.

FIGURE 17.8

Excel slicers for a PivotTable based on the Chapter 2 sample of 316 retirement funds

Slicers are not included in OS X Excel 2011 and Microsoft Windows versions of Excel that are older than Excel 2010



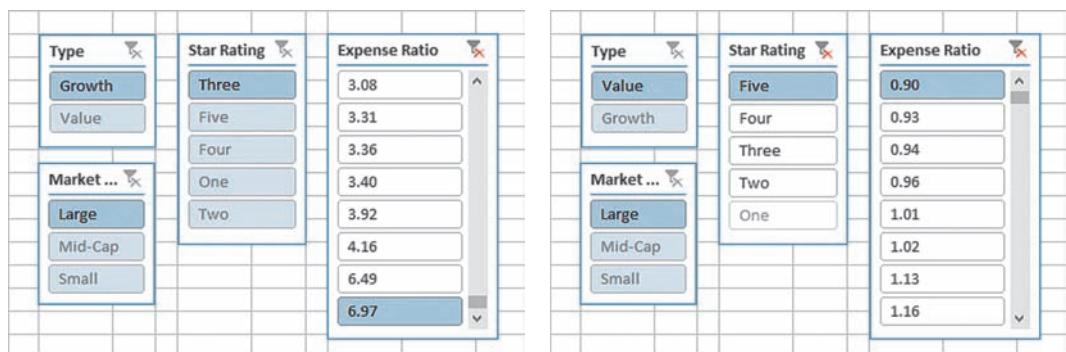
Slicers allow you to work with more than three or four variables at the same time in a way that avoids creating overly complex multidimensional contingency tables that would be hard to read. By clicking specific buttons in slicer panels, you can filter data to ask questions of the data you have collected. For example, with the Figure 17.8 slicer panels you could ask questions such as

1. What are the attributes of the fund with the highest expense ratio?
2. What is the type and market cap of the five-star fund with the lowest expense ratio?

Figure 17.9 shows slicer displays that answer these two questions. Note that Excel has disabled, or dimmed, the buttons that represent values that the current filtering excludes. This allows you to visually discover the answers to questions. For example, the answer to Question 1 is a large market cap growth fund that is rated three stars. For Question 2, the answer is a large market cap value fund.

FIGURE 17.9

Slicer displays for Questions 1 (left) and 2 (right)



Because data discovery provides a means of preliminary analysis, data discovery is often part of many predictive analytics methods. Later in this chapter, you will recognize data discovery techniques that are embedded in cluster analysis and multidimensional scaling.

Problems for Section 17.1

17.1 The Edmunds.com NHTSA Complaints Activity Report is the result of the examination of the frequency, trends and composition of consumer vehicle complaint submissions at the automaker, brand, and category levels (data extracted from [edmu.in/Ybmpuz](#)). The table at the right, stored in **Automaker1**, contains complaints received by six automakers for January 2013. When the number of complaints are below 300, the complaint rating is considered to be low, when the number of complaints is between 300 and 500, the complaint rating is considered to be medium, and when the number of complaints is above 500, the complaint rating is considered to be high.

Automaker	Number of Complaints
American Honda	169
Chrysler LLC	439
Ford Motor Company	440
General Motors	551
Nissan Motors Corporation	467
Toyota Motor Sales	332

- a. Construct a gauge for each automaker.
- b. Construct a bullet graph for the automakers.
- c. Which display is more effective at comparing the number of complaints for each automaker?

17.2 There are a very large number of mutual funds from which an investor can choose. Each mutual fund has its own mix of different types of investments. The file **BestFunds1** contains the one-year return percentage and the three-year annualized return percentage for the ten best short-term bond and long-term bond funds according to the *U.S. News & World Report* score. (Data extracted from money.usnews.com/mutual-funds/rankings.)

- Construct bullet graphs of the one-year returns and the three-year returns. For the purposes of comparison, consider a return below 5% as low-performing, a return between 5 and 10% as medium-performing, and a return above 10% as high-performing.
- Why would you not want to construct a gauge for each bond fund?
- What conclusions can you reach about the one-year and three-year return percentages for the short-term bond and long-term bond funds?

17.3 A financial analyst was interested in comparing the price-to-book ratio (P/B) of drug companies. The analyst collected P/B ratios for 71 drug companies (Industry Group SIC 3 code: 283) and stored them as part of the file **BusinessValuation**.

- Visually evaluate the P/B ratios by constructing a bullet graph. For the purposes of comparison, consider a P/B ratio that is 2 or less as excellent, a P/B ratio that is between 2 and 5 as acceptable, and a P/B ratio that is above 5 as unacceptable.
- Why would using gauges be a poor choice for this analysis?
- Are the three groupings of P/B ratios helpful in analyzing the data? What constitutes an acceptable P/B ratio varies by industry and is partially based on subjective analysis. For the purposes of information presentation, would you redefine or subdivide the current acceptable category?

17.4 The file **BBCost2012** contains the total cost (in \$) for four tickets, two beers, four soft drinks, four hot dogs, two game programs, two baseball caps, and parking for one vehicle at each of the 30 Major League Baseball (MLB) parks during the 2012 season. (Data extracted from fancostexperience.com.)

- Visually evaluate the total cost at each MLB park by constructing a bullet graph. For the purposes of comparison, consider a total cost (in \$) below \$180 as inexpensive, between \$180 and \$240 as typical, and above \$240 as expensive.
- Compare the bullet graph constructed in (a) with the stem-and-leaf display constructed in Problem 2.36 on page 63 for the same data.
- Which display best visualizes the distribution of costs? Why?
- Name something that the bullet graph reveals about the data that the stem-and-leaf display does not. How could that be used as the basis for future analysis of total costs at MLB parks?

17.5 Referring to the data in Problem 2.56 on page 68 concerning movie attendance between 2002 and 2012 (stored in **MovieAttendance2**):

- Construct a sparkline graph for the movie attendance between 2002 and 2012.
- What conclusions can you reach about the movie attendance between 2002 and 2012?
- Compare the sparkline graph with the time-series plot constructed in Problem 2.56 on page 68 for the same data. When would using the sparkline graph be the better choice to visualize

these data? When would using the time-series plot be the better choice?

- Might you ever use both a sparkline graph and a time-series plot in the same analysis report? Explain your reasoning.

17.6 The file **StockIndices** contains the data that represent the total rate of return (in percentage) for the Dow Jones Industrial Average (DJIA), the Standard & Poor's 500 (S&P 500), and the technology-heavy NASDAQ Composite (NASDAQ) from 2006 through 2012. Source: Data extracted from finance.yahoo.com, March 29, 2013.

- Construct sparklines for the rate of return per year for the DJIA, S&P 500, and NASDAQ from 2006 through 2012.
- What conclusions can you reach concerning the rates of return per year of the three market indices?
- Compare the results of (b) to those of Problem 17.7 (b).

17.7 From 2006 to 2012, the value of precious metals fluctuated dramatically. The file **Metal Indices** contains the total rate of return (in percentage) for platinum, gold, and silver from 2006 through 2012. (Data extracted from finance.yahoo.com, March 29, 2013.)

- Construct sparklines for rate of return per year for platinum, gold, and silver from 2006 through 2012.
- What conclusions can you reach concerning the rates of return of the three precious metals?
- Compare the results of (b) to those of Problem 17.6 (b).

17.8 Drive-through service time is an important quality attribute for fast food chains. The data in **ServiceTime** are the mean service times for Burger King, Chick-Fil-A, McDonald's, and Wendy's in 12 recent years. (Data extracted from bit.ly/qhvP3Zb.)

- Construct sparklines of the mean service times for Burger King, Chick-Fil-A, McDonald's, and Wendy's in 12 recent years.
- What conclusions can you reach concerning the mean service times for Burger King, Chick-Fil-A, McDonald's, and Wendy's in 12 recent years?

17.9 Sales of automobiles in the United States fluctuate from month to month and year to year. The data in the file **AutoSales** represent the sales for various manufacturers in July 2013 and the change from July 2012 sales in percentages. (Data extracted from www.nytimes.com/interactive/2013/08/01/business/How-the-Auto-Industry-Fared-in-July.htm.)

- Construct a treemap of the sales of autos and the change in sales from July 2012.
- What conclusions can you reach concerning the sales of autos and the change in sales from July 2012?

17.10 The value of a National Basketball Association (NBA) franchise has increased dramatically over the past few years. The value of a franchise varies based on the size of the city in which the team is located, the amount of revenue it receives, and the success of the team. The file **NBAValues** contains the value of each team and the change in value in the past year. (Data extracted from www.forbes.com/nba-valuations.)

- Construct a treemap that visualizes the values of the NBA teams (size) and the one year changes in value (color).
- What conclusions can you reach concerning the value of NBA teams and the one year change in value?

17.11 The annual ranking of the FT Global 500 2013 provides a snapshot of the world's largest companies. The companies are ranked by market capitalization—the greater the stock market value of a company, the higher the ranking. The market capitalizations (in \$billions) and the 52-week change in market capitalizations (in percentages) for companies in the Automobile & Parts, Financial Services, Health Care Equipment & Services, and Software & Computer Services sectors are stored in **FTGlobal500**. (Data extracted from ft.com/intl/indepth/ft500.)

- Construct a treemap that presents each company's market capitalization (size) and the 52-week change in market capitalization (color) grouped by sector and country.
- Which sector seems to have the best gains in the market capitalizations of its companies? Which sectors seem to have the worst gains (or greatest losses)?
- Construct a treemap that presents each company's market capitalization (size) and the 52-week change in market capitalization (color) grouped by country.
- What comparison can be more easily made with the treemap constructed in (c) compared to the treemap constructed in (a)?

17.12 Your task as a member of the International Strategic Management Team at your company is to investigate the potential for entry into a foreign market. As part of your initial investigation, you must provide an assessment of the economies of countries in the Americas and the Asia and Pacific regions. The file **DoingBusiness** contains the 2012 GDPs per capita for these countries as well as the number of Internet users in 2011 (per 100 people) and the number of mobile cellular subscriptions in 2011 (per 100 people). (Data extracted from data.worldbank.org.)

- Construct a treemap of the GDPs per capita (size) and their number of Internet users in 2011 (per 100 people) (color) for each country grouped by region.
- Construct a treemap of the GDPs per capita (size) and their number of mobile cellular subscriptions in 2011 (per 100 people) (color) for each country grouped by region.
- What patterns do these data suggest? Are the patterns in the two treemaps similar or different? Explain.

17.13 Using the sample of retirement funds stored in **Retirement Funds**:

- Construct a table that tallies type, market cap, and risk.
- Drill down to examine the large cap growth funds with high risk. How many funds are there? What conclusions can you reach about these funds?

17.14 Using the sample of retirement funds stored in **Retirement Funds**:

- Construct a table that tallies type, market cap, and rating.
- Drill down to examine the large cap growth funds with a rating of three. How many funds are there? What conclusions can you reach about these funds?

17.15 Using the sample of retirement funds stored in **Retirement Funds**:

- Construct a table that tallies market cap, risk, and rating.
- Drill down to examine the large cap funds that are high risk with a rating of three. How many funds are there? What conclusions can you reach about these funds?

17.16 Using the sample of retirement funds stored in **Retirement Funds**:

- Construct a table that tallies type, risk, and rating.
- Drill down to examine the growth funds with high risk with a rating of three. How many funds are there? What conclusions can you reach about these funds?

17.17 Using the sample of retirement funds stored in **Retirement Funds**, what are the attributes of the fund with the highest five-year return?

17.18 Using the sample of retirement funds stored in **Retirement Funds**, what five-year returns are associated with small market cap funds that have a rating of five stars?

17.19 Using the sample of retirement funds stored in **Retirement Funds**, which fund(s) in the sample have the lowest five-year return?

17.20 Using the sample of retirement funds stored in **Retirement Funds**, what is the type and market cap of the five-star fund with the highest five-year return?

17.2 Predictive Analytics

Predictive analytics are methods that determine what is likely to occur in the future. These methods fall into one of four categories:

- Prediction: Assigning a value to a target based on a model
- Classification: Assigning items in a collection to target categories or classes
- Clustering: Finding natural groupings in data
- Association: Finding items that tend to co-occur and specifying the rules that govern their co-occurrence

Earlier chapters have already discussed some prediction and classification methods including regression and correlation and logistic regression. This chapter discusses classification and regression trees, neural networks, cluster analysis, and multidimensional scaling.

While predictive analytics has been used with samples from a population, the growing use of *big data* has led to more widespread use. Combining predictive methods with big data requires *data mining*. **Data mining** are the techniques that allow the extraction of

useful knowledge from huge repositories of data. That extraction is rarely simple and typically requires statistical methods and computer science algorithms as well as the database-type operations associated with simple searches and retrievals of data. Because of this common association of predictive analytics and data mining, some people intertwine the meanings of the two terms or use them jointly in the phrase “predictive analytics and data mining.” Even though predictive analytics methods may be *most* useful to a business decision maker when used with data mining, these methods can exist independently of data mining and can be used with smaller sets of data, as the examples used in this chapter illustrate.

Table 17.1 identifies and classifies the specific predictive analytics methods that are discussed in the remainder of this chapter.

TABLE 17.1

Chapter 17 Predictive Analytics Methods

METHOD	METHOD FOR			
	Prediction	Classification	Clustering	Association
Classification and Regression Trees	•	•		
Neural Networks	•	•	•	
Cluster Analysis			•	
Multidimensional Scaling (MDS)		•		•

17.3 Classification and Regression Trees

Section 4.2 introduced decision trees, a visual alternative to contingency tables. **Classification and regression trees** are decision trees that split data into groups based on the values of independent or explanatory (X) variables. This splitting determines which values of a specific independent variable are useful in predicting the dependent (Y) variable.

Using a *categorical* dependent Y variable results in a *classification tree*. Using a *numerical* dependent Y variable results in a *regression tree*. To construct the tree, the classification and regression tree method chooses the independent variable that provides the best split of the data at each node in the tree, starting with the root. Successfully completing a classification or regression tree requires:

- Rules for splitting the data at each node based on one independent variable.
- Rules for deciding when a branch cannot be split any more.
- A prediction for the target variable at each node.

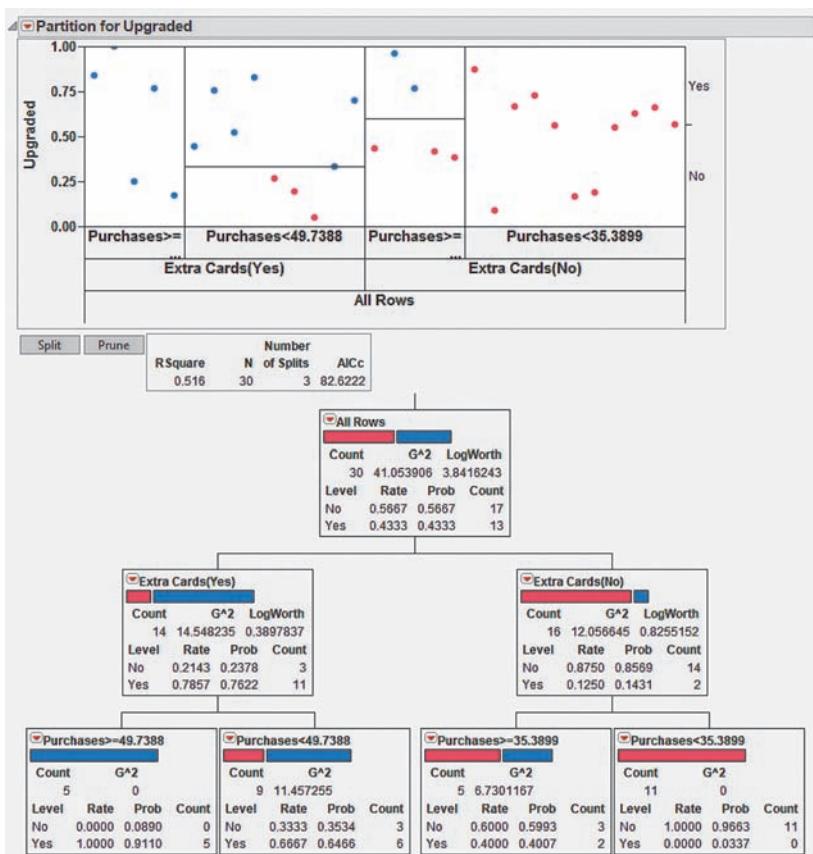
When using classification and regression trees, the variables included in the model are not determined prior to the analysis but are selected based on the algorithm used by the classification and regression tree process. For example, the statistical software JMP uses the **Gini impurity**, the product of the probability that each item is chosen multiplied by the probability that the item has been misclassified, as a basis to determine how to split items at each node (see reference 6).

Classification and regression trees are not affected by the distribution of the variables that make up the data. Typically, trees are developed through several layers of branches until either no further gain in the fit occurs or the splitting has continued as far as possible. Splitting can be followed by pruning back the tree as is necessary. As with regression modeling, in order to validate any classification and regression tree analysis, *where possible*, you should split the data into a training sample that is used to develop models for analysis and a validation sample that is used to determine the validity of the models developed in the training sample.

To illustrate a classification tree analysis, return to the Section 14.7 card study example on page 574 in which a logistic regression model was used to predict the proportion of credit card holders who would upgrade to a premium card. Using JMP, you can create the classification tree results shown in Figure 17.10.

FIGURE 17.10

Classification tree results for predicting the proportion of credit card holders who would upgrade to a premium card



In the tree diagram portion of Figure 17.10, the root represents all the data. The Yes rate, 0.4333, reflects the 13 out of 30 cardholders who have upgraded to a premium card. (The No rate, 0.5667, reflects the 17 cardholders who have not upgraded.) The first split of the data is based on whether the cardholder has additional cards. In the Extra Cards(Yes) node that represents those cardholders who have extra cards, the Yes rate, 0.7857, represents the 11 of these 14 who have upgraded. In the Extra Cards(No) node that represents those cardholders who do not have extra cards, the Yes rate, 0.125, represents the 2 out of 16 who have upgraded.

At the next level, the two nodes are split based on the amount of annual purchases. For the Extra Cards (Yes) node, the split is between those who charge more than \$49,738.80 per year and those who charge less than \$49,738.80 per year. For the Extra Cards(No) node, the split is between those who charge more than \$35,389.90 per year and those who charge less than \$35,389.90 per year. For those cardholders who have extra cards and use the card to purchase more than \$49,738.80 per year (the left node on the bottom level), the rate of 1.0 reflects that all five have upgraded. For those cardholders who have extra cards and use the card to purchase less than \$49,738.80 per year (second from left), the rate 0.6667 reflects that 6 of the 9 have upgraded. For those cardholders who do not have extra cards and use the card to purchase more than \$35,389.90 per year (second from right), the rate 0.4000 reflects that 2 of the 5 have upgraded. For those cardholders who do not have extra cards and use the card to purchase less than \$35,389.90 per year (right node on the bottom level), the rate of 0 reflects that none of the 11 have upgraded.

These results show that customers who order extra cards are *more* likely to upgrade a premium card and that those customers in this group who charge extensively are *most* likely to upgrade. (Least likely to upgrade to a premium card are customers who have only a single charge card and have charged less than \$35,389.90.) Therefore, managers at the credit card company

might want to make a special appeal to customers who charge extensively and have ordered extra cards while creating a broader marketing effort aimed at all those who have ordered extra cards.

The r^2 of 0.516, shown in the summary box below the plot, means that 51.6% of the variation in whether a cardholder upgrades can be explained by the variation in whether the cardholder has additional cards and the amount the cardholder charges per year. The **Akaike information criterion (AIC)** and the AIC_c , the “corrected” version of the AIC, measure the relative quality of the model.

AKAIKE INFORMATION CRITERION (AIC)

$$AIC = 2k - 2 \ln(L) \quad (17.1)$$

where

k = the number of parameters in the model

L is the maximum value of the likelihood function for the model

AKAIKE INFORMATION CRITERION CORRECTED (AIC_c)

AIC_c corrects AIC for the sample size.

$$AIC_c = AIC + \frac{2k(k + 1)}{n - k - 1} \quad (17.2)$$

where

n = sample size

Using the AIC_c provides a way of comparing alternative models in terms of information loss. At each branch, JMP reports the LogWorth and G^2 statistics. The **LogWorth statistic** is the basis for splitting at each node of the classification tree. LogWorth is calculated as:

LOGWORTH

$$\text{LogWorth} = -\log_{10}(p\text{-value}) \quad (17.3)$$

where the adjusted p -value is based on the number of ways that splits can occur.

The G^2 statistic is the likelihood ratio chi-square statistic. It is a measure of the probability that can be attributed to the response that has occurred. In Figure 17.10, the **Prob** column is the predicted probability for that node of the tree.

Student Tip

Remember that a regression tree results from using a *numerical* dependent variable Y .

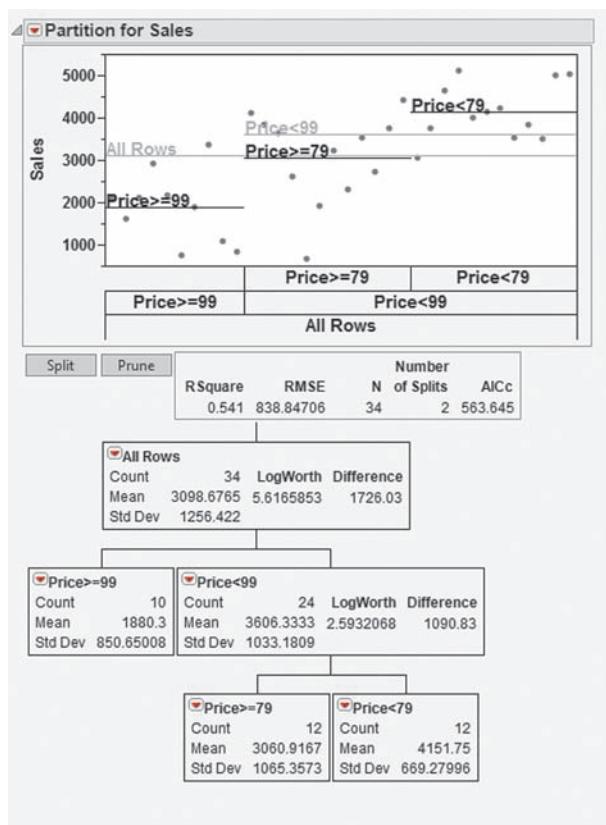
Regression Tree Example

To illustrate a regression tree analysis, return to the Chapter 14 OmniPower scenario. As discussed on page 543, you are a marketing manager to determine the effect that price and in-store promotional expenses will have on sales of OmniPower bars.

Figure 17.11 on page 686 presents the regression tree constructed by JMP for predicting the sales of OmniPower bars.

FIGURE 17.11

Regression tree results for predicting the sales of OmniPower bars



Observe from Figure 17.11 that the tree root represents all the data. The r^2 value of 0.541 means that 54.1% of the variation in sales can be explained by variation in the price and promotional expenses.

The first split in the tree is based on the price being less than 99 cents or greater than or equal to 99 cents. For the 10 stores in which the price was greater than or equal to 99 cents, the mean sales were \$1,880. For the 24 stores in which the price was less than 99 cents, the mean sales were \$3,606.3333.

The next split is on the portion of the tree that contained the 24 stores in which the price was less than 99 cents. For the 12 stores in which the price was greater than or equal to 79 cents, the mean sales were \$3,060.9167. For the 12 stores in which the price was less than 79 cents, the mean sales were \$4,151.75.

You can conclude that sales were much higher at stores in which the price was below 99 cents. (Because the prices were set as either 59, 79, or 99 cents, this result means that sales were higher at stores where the bars were sold at 59 or 79 cents.) Going one step further, you can conclude that sales were higher at stores where the price of OmniPower bars was below 79 cents (in other words, 59 cents). Given these results, the OmniFoods marketing manager could then determine how pricing the bars at 59 cents would impact their profitability.

Problems for Section 17.3

17.21 A hotel has designed a new system for room service delivery of breakfast that allows the customer to select a specific delivery time. The file **Satisfaction** contains the difference between the actual and requested delivery times (a negative time means that the breakfast was delivered before the requested time) recorded for 30 deliveries on a particular day along with whether the customer had previously stayed at the hotel.

- Using all the data as the training sample, develop a classification tree model to predict the probability that the customer will be satisfied based on the delivery time difference and whether the customer had previously stayed at the hotel.
- What conclusions can you reach about the probability that the customer will be satisfied?

17.22 A marketing manager wants to predict customers with risk of churning (switching their service contracts to another company) based on the number of calls the customer makes to the company call center and the number of visits the customer makes to the local service center. Data from a random sample of 30 customers are organized and stored in **Churn**.

- Using all the data as the training sample, develop a classification tree model to predict the probability of churning, based on the number of calls the customer makes to the company call center and the number of visits the customer makes to the local service center.
- What conclusions can you reach about the probability of churning?

17.23 An automotive insurance company wants to predict which filed stolen vehicle claims are fraudulent, based on the number of claims submitted per year by the policy holder and whether the policy is a new policy, that is, is one year old or less (coded as 1 = yes, 0 = no). Data from a random sample of 98 automotive insurance claims are organized and stored in **InsuranceFraud**. (Data extracted from Gelp et al., “A Comparative Analysis of Decision Trees vis-à-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection,” *Journal of Data Science*, 10 (2012), pp. 537–561.)

- Using all the data as the training sample, develop a classification tree model to predict the probability of a fraudulent claim, based on the number of claims submitted per year by the policy holder and whether the policy is new.
- What conclusions can you reach about the probability of a fraudulent claim?
- Using half the data as the training sample and the other half of the data as the validation sample, develop a classification tree model to predict the probability of a fraudulent claim, based on the number of claims submitted per year by the policy holder and whether the policy is new.
- What differences exist in the results of (a) and (c)? What conclusions can you reach about the models fit from the training samples in (a) and (c)?

17.24 Undergraduate students at Miami University in Oxford, Ohio, were surveyed in order to evaluate the effect of price on the purchase of a pizza from Pizza Hut. The students were asked to suppose that they were going to have a large two-topping pizza delivered to their residence. Then they were asked to select from either Pizza Hut or another pizzeria of their choice. The price they would have to pay to get a Pizza Hut pizza differed from survey to survey. For example, some surveys used the price \$11.49. Other prices investigated were \$8.49, \$9.49, \$10.49, \$12.49, \$13.49, and \$14.49. The dependent variable for this study is whether or not a student will select Pizza Hut. The independent variables are the price of a Pizza Hut pizza and the gender of the student (1 = male, 0 = female). The results of these surveys are stored in **PizzaHut**.

- Using half the data as the training sample and the other half of the data as the validation sample, develop a classification tree model to predict the probability the student will select Pizza Hut based on the price of a Pizza Hut pizza and the gender of the student.
- What conclusions can you reach about the probability the student will select Pizza Hut?

17.25 The business problem facing a consumer products company is to measure the effectiveness of different types of advertising media in the promotion of its products. Specifically, the company is interested in the effectiveness of radio advertising and newspaper advertising (including the cost of discount coupons). During a one-month test period, data were collected from a sample of 22 cities with approximately equal populations. Each city is allocated a specific expenditure level for radio advertising and for newspaper advertising. The sales of the product (in thousands of dollars) and also the levels of media expenditure (in thousands of dollars) during the test month are recorded and stored in **Advertise**.

- Using all the data as the training sample, develop a regression tree model to predict the sales of the product.
- What conclusions can you reach about the sales of the product?

17.26 Starbucks Coffee Co. uses a data-based approach for improving the quality and customer satisfaction of its products. When survey data indicated that Starbucks needed to improve its package sealing process, an experiment was conducted (data extracted from L. Johnson and S. Burrows, “For Starbucks, It’s in the Bag,” *Quality Progress*, March 2011, pp. 17–23) to determine the factors in the bag-sealing equipment that might be affecting the ease of opening the bag without tearing the inner liner of the bag. Among the factors that could affect the rating of the ability of the bag to resist tears were the viscosity, pressure, and plate gap on the bag-sealing equipment. Data were collected on 19 bags in which the plate gap was varied. The results are stored in **Starbucks**.

- Using all the data as the training sample, develop a regression tree model to predict the rating of the ability of the bag to resist tears.
- What conclusions can you reach about the rating of the ability of the bag to resist tears?

17.27 In mining engineering, holes are often drilled through rock using drill bits. As a drill hole gets deeper, additional rods are added to the drill bit to enable additional drilling to take place. It is expected that drilling time increases with depth. This increased drilling time could be caused by several factors, including the mass of the drill rods that are strung together. The business problem relates to whether drilling is faster using dry drilling holes or wet drilling holes. Using dry drilling holes involves forcing compressed air down the drill rods to flush the cuttings and drive the hammer. Using wet drilling holes involves forcing water rather than air down the hole. Data have been collected from a sample of 50 drill holes that contains measurements of the time to drill each additional 5 feet (in minutes), the depth (in feet), and whether the hole was a dry drilling hole or a wet drilling hole. The data are organized and stored in **Drill**.

- Using half the data as the training sample and the other half of the data as the validation sample, develop a regression tree model to predict the drilling time.
- What conclusions can you reach about the drilling time?

17.28 The owner of a moving company typically has his most experienced manager predict the total number of labor hours that will be required to complete an upcoming move. This approach has proved useful in the past, but the owner has the business objective of developing a more accurate method of predicting labor hours. In a preliminary effort to provide a more accurate method,

the owner has decided to use the number of cubic feet moved, the number of large pieces of furniture, and whether there is an elevator in the apartment building as the independent variables and has collected data for 36 moves in which the origin and destination were within the borough of Manhattan in New York City and the

travel time was an insignificant portion of the hours worked. The data are organized and stored in **Moving**.

- Using all the data as the training sample, develop a regression tree model to predict the labor hours.
- What conclusions can you reach about the labor hours?

17.4 Neural Networks

³Because neural networks were inspired by the architecture of the human brain, some describe such networks as *learning* from the data to form the best model. This “learning” in neural networks, while complex, is more analogous to the computation and evaluation of inferential methods discussed earlier in this book than to the many types of human learning. Therefore this book uses “uncovers” where other sources might use “learn.”

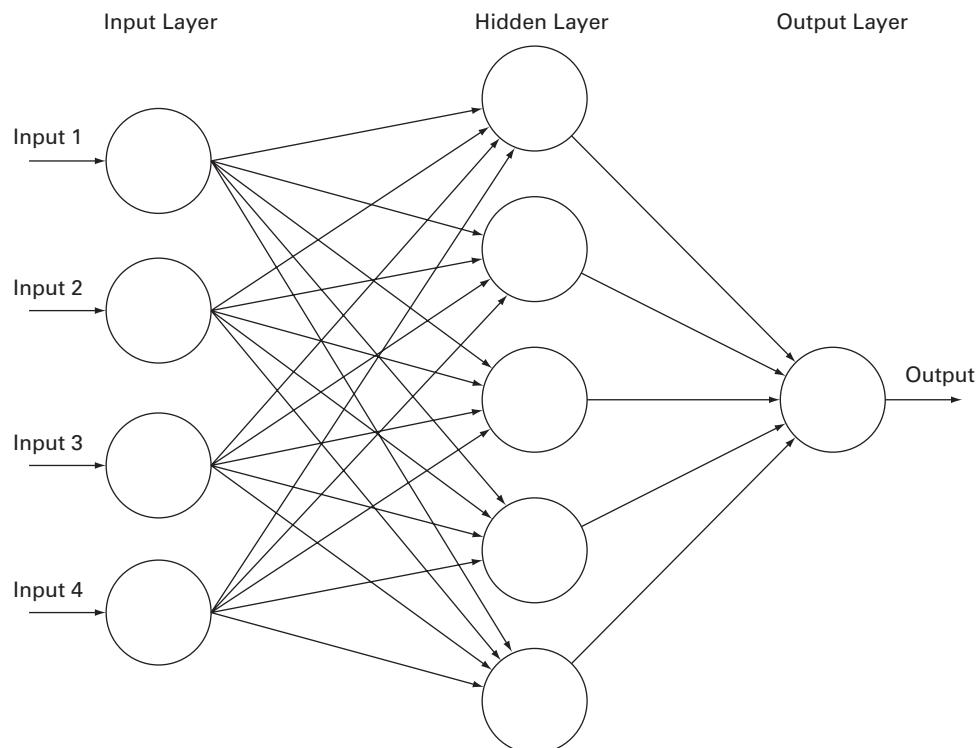
Neural networks are powerful, flexible data mining techniques that construct models from patterns and relationships uncovered in data. Unlike the inferential methods in earlier chapters, in which you must supply a model to be tested, neural networks “learn” from the data to construct that model for you.³ Neural networks are very flexible and can be applied to prediction, classification, and clustering problems and, as nonparametric methods, do not make *a priori* assumptions about the distribution of the data. Because they do not require a supplied model as a starting point, neural networks are particularly valuable in analyzing big data in which the process of model transformations discussed in Chapters 14 and 15 would be too unwieldy and time consuming to perform.

Multilayer Perceptrons

All neural networks contain complex computations that begin with *inputs* and end with *outputs*. Neural networks used for prediction and classification are typically **multilayer perceptrons (MLPs)** that contain an *input layer*, a *hidden layer*, and an *output layer*, as shown in Figure 17.12.

FIGURE 17.12

Structure of a multilayer perceptron



⁴Processing elements are meant to simulate the neuron in the brain and therefore are often referred to as *artificial neurons* or *nodes*. For more on the relationship between biological and artificial neural networks, see reference 10.

To construct models, MLPs use *error back propagation*. To begin, the input layer nodes (the circles in Figure 17.12) send the various inputs to the nodes of the hidden layer. The inputs have associated weights. **Processing elements**⁴ that comprise the hidden layer combine the weighted inputs and apply a nonlinear function, such as the S-shaped hyperbolic tangent function, to the combination. The **hyperbolic tangent function** is a function that varies between -1 and $+1$.

HYPERBOLIC TANGENT FUNCTION

$$\frac{e^{2x} - 1}{e^{2x} + 1} \quad (17.4)$$

where x is a linear combination of the X variables.

LEARN MORE

Learn more about the mathematical computations used in MLP neural networks in the SHORT TAKES for Chapter 17.

The results of the processing element computations in the hidden layer are sent to the output layer. The output layer combines the results it receives from the hidden layer and compares it to the target Y value. The output layer then sends back to the hidden layer nodes (the start of the *back propagation*) its estimate of the difference between the predicted results and the target value (the error rate). Computations in the hidden layer then backwardly influence the weighting done near the input layer, and the process continues forward a second time to the output layer. This forward-and-backward computation among the three layers continues until the output layer detects that the error rate has been minimized or is at an acceptable level. (This minimization of errors is analogous to the minimization of prediction error of regression models.) At this point, the model is established.

To establish a neural network, you use some of your data as the *training data* and some of it as the *validation data*. Neural networks use the **training data** to uncover a model that by some criteria best describes the patterns and relationships in the data. The model is then applied to the *validation data* to see if the model can make the correct prediction or classification.

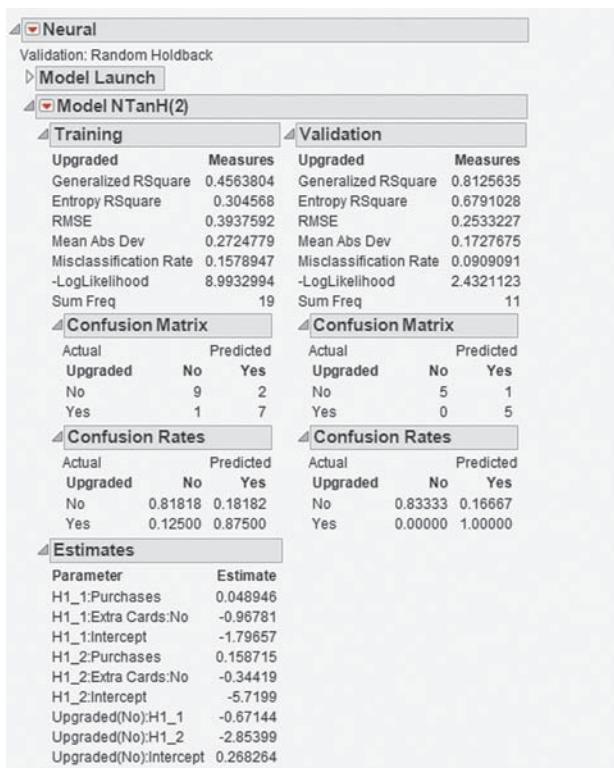
Models that neural networks construct can be difficult to interpret (see reference 5). Neural networks also can suffer from poor quality of data, insufficient data, or *overfitted* models, models that only work well with the data used to construct that model. Determining the number of hidden nodes (processing elements) can be inexact process. One source suggests that when using a neural network for classification to start with one hidden node per class (see reference 7).

To illustrate an MLP, recall the Section 14.7 card study example on page 574 that sought to identify credit cardholders who would be likely to upgrade to a premium card. Figure 17.13 shows the MLP results computed by JMP for both the training and validation samples.

FIGURE 17.13

Multilayer Perceptron results for classifying credit card holders who would upgrade to a premium card

The table of parameter estimates of the model contains the final estimated weights that resulted from the backpropagation training process. Even though these weights cannot be interpreted in the same manner as regression coefficients, weights are crucial in that they store the patterns that were uncovered (“learned”) in analyzing the data.

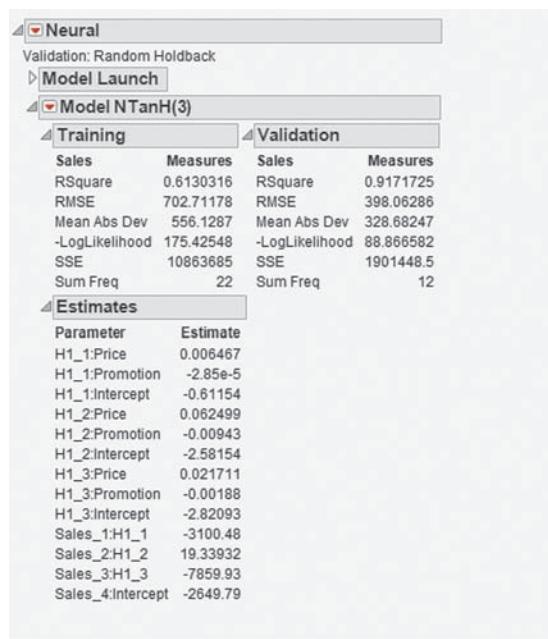


The validation data results can be used as a representation of the model's predictive power on future observations. The validation data misclassification rate, 0.0909, means that 9% of the cardholders in the validation set were inaccurately classified using the trained MLP neural network. The Confusion Matrix report shows a contingency table of the actual and predicted values for the Upgraded variable. The Confusion Rates report is equal to the Confusion Matrix report, with the numbers divided by the row totals. Of the 11 cardholders in the validation set, 5 actually upgraded to the premium card and 6 did not. Of the 5 cardholders who actually did upgrade to the premium card, 5 or 1.0000 were correctly classified by the MLP neural network; of the 6 cardholders who actually did not upgrade to the premium card, 5 or 0.8333 were correctly classified by the MLP NN.

To illustrate a neural network analysis with a continuous numerical dependent variable, return to the Chapter 14 OmniPower scenario on page 543. Figure 17.14 presents the neural network results for predicting the sales of OmniPower bars. The r^2 statistic for the validation data is 0.9172. This indicates that the model is doing a good job in predicting the target in the validation data.

FIGURE 17.14

Multilayer perceptron results for predicting the sales of OmniPower bars



Problems for Section 17.4

17.29 A hotel has designed a new system for room service delivery of breakfast that allows the customer to select a specific delivery time. The file **Satisfaction** contains the difference between the actual and requested delivery times recorded (a negative time means that the breakfast was delivered before the requested time) for 30 deliveries on a particular day along with whether the customer had previously stayed at the hotel.

- Develop a neural net model to predict the probability that the customer will be satisfied, based on the delivery time difference and whether the customer had previously stayed at the hotel.
- What conclusions can you reach about the probability that the customer will be satisfied?

17.30 A marketing manager wants to predict customers with risk of churning (switching their service contracts to another company) based on the number of calls the customer makes to the company call center and the number of visits the customer makes to the local service center. Data from a random sample of 30 customers are organized and stored in **Churn**.

- Develop a neural net model to predict the probability of churning, based on the number of calls the customer makes to the

company call center and the number of visits the customer makes to the local service center.

- What conclusions can you reach about the probability of churning?

17.31 An automotive insurance company wants to predict which filed stolen vehicle claims are fraudulent, based on the number of claims submitted per year by the policy holder and whether the policy is a new policy, that is, is one year old or less (coded as 1 = yes, 0 = no). Data from a random sample of 98 automotive insurance claims are organized and stored in **InsuranceFraud**. (Data extracted from Gelp et al., "A Comparative Analysis of Decision Trees vis-à-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection," *Journal of Data Science*, 10 (2012), pp. 537–561.)

- Develop a neural net model to predict the probability of a fraudulent claim, based on the number of claims submitted per year by the policy holder and whether the policy is new.
- What conclusions can you reach about the probability of a fraudulent claim?

17.32 Undergraduate students at Miami University in Oxford, Ohio, were surveyed in order to evaluate the effect of price on the purchase of a pizza from Pizza Hut. The students were asked to suppose that they were going to have a large two-topping pizza delivered to their residence. Then they were asked to select from either Pizza Hut or another pizzeria of their choice. The price they would have to pay to get a Pizza Hut pizza differed from survey to survey. For example, some surveys used the price \$11.49. Other prices investigated were \$8.49, \$9.49, \$10.49, \$12.49, \$13.49, and \$14.49. The dependent variable for this study is whether or not a student will select Pizza Hut. The independent variables are the price of a Pizza Hut pizza and the gender of the student (1 = male, 0 = female). The results of these surveys are stored in [PizzaHut](#).

- Develop a neural net model to predict the probability the student will select Pizza Hut based on the price of a Pizza Hut pizza and the gender of the student.
- What conclusions can you reach about the probability the student will select Pizza Hut?

17.33 The business problem facing a consumer products company is to measure the effectiveness of different types of advertising media in the promotion of its products. Specifically, the company is interested in the effectiveness of radio advertising and newspaper advertising (including the cost of discount coupons). During a one-month test period, data were collected from a sample of 22 cities with approximately equal populations. Each city is allocated a specific expenditure level for radio advertising and for newspaper advertising. The sales of the product (in thousands of dollars) and also the levels of media expenditure (in thousands of dollars) during the test month are recorded and stored in [Advertise](#).

- Develop a neural net model to predict the sales of the product.
- What conclusions can you reach about the sales of the product?

17.34 Starbucks Coffee Co. uses a data-based approach for improving the quality and customer satisfaction of its products. When survey data indicated that Starbucks needed to improve its package sealing process, an experiment was conducted (data extracted from L. Johnson and S. Burrows, "For Starbucks, It's in the Bag," *Quality Progress*, March 2011, pp. 17–23) to determine the factors in the bag-sealing equipment that might be affecting the ease of opening the bag without tearing the inner liner of the bag. Among the factors that could affect the rating of the ability of the bag to

resist tears were the viscosity, pressure, and plate gap on the bag-sealing equipment. Data was collected on 19 bags in which the plate gap was varied. The results are stored in [Starbucks](#).

- Develop a neural net model to predict the rating of the ability of the bag to resist tears.
- What conclusions can you reach about the rating of the ability of the bag to resist tears?

17.35 In mining engineering, holes are often drilled through rock, using drill bits. As a drill hole gets deeper, additional rods are added to the drill bit to enable additional drilling to take place. It is expected that drilling time increases with depth. This increased drilling time could be caused by several factors, including the mass of the drill rods that are strung together. The business problem relates to whether drilling is faster using dry drilling holes or wet drilling holes. Using dry drilling holes involves forcing compressed air down the drill rods to flush the cuttings and drive the hammer. Using wet drilling holes involves forcing water rather than air down the hole. Data have been collected from a sample of 50 drill holes and contain measurements of the time to drill each additional 5 feet (in minutes), the depth (in feet), and whether the hole was a dry drilling hole or a wet drilling hole. The data are organized and stored in [Drill](#).

- Develop a neural net model to predict the drilling time.
- What conclusions can you reach about the drilling time?

17.36 The owner of a moving company typically has his most experienced manager predict the total number of labor hours that will be required to complete an upcoming move. This approach has proved useful in the past, but the owner has the business objective of developing a more accurate method of predicting labor hours. In a preliminary effort to provide a more accurate method, the owner has decided to use the number of cubic feet moved, the number of large pieces of furniture, and whether there is an elevator in the apartment building as the independent variables and has collected data for 36 moves in which the origin and destination were within the borough of Manhattan in New York City and the travel time was an insignificant portion of the hours worked. The data are organized and stored in [Moving](#).

- Develop a neural net model to predict the labor hours.
- What conclusions can you reach about the labor hours?

17.5 Cluster Analysis

Cluster analysis seeks to classify data into a sequence of groupings such that objects in each group are more alike other objects in their group than they are to objects found in other groups. Cluster analysis can be performed in several different ways; two of the most common techniques are *hierarchical clustering* and *k-means clustering*.

In **hierarchical clustering**, the analysis starts with each object in its own cluster. Then, the two objects that are determined to be the closest to each other are merged into a single cluster. The merging of the two closest objects repeats until there remains only one cluster that includes all objects. In **k-means clustering**, the number of clusters (k) is set at the start of the process. Objects are then assigned to clusters in an iterative process that seeks to make the means of the k clusters as different as possible. During the iterative process, unlike hierarchical clustering, in which clusters once formed are never changed later in the process, objects may be reassigned to a different cluster later in the process.

To perform a cluster analysis, you must determine how to measure the distance between objects and how to measure the distance between clusters. The most common measure of

Student Tip

Different clustering techniques can produce different results

distance between objects used in cluster analysis is **Euclidean distance** that measures the distance between objects as the square root of the sum of the squared differences between objects over all r dimensions.

EUCLIDEAN DISTANCE (CLUSTER ANALYSIS)

$$d_{ij} = \sqrt{\sum_{k=1}^r (X_{ik} - X_{jk})^2} \quad (17.5)$$

where

d_{ij} = distance between object i and object j

X_{ik} = value of object i in dimension k

X_{jk} = value of object j in dimension k

r = number of dimensions

Among the measures of distance between clusters are complete linkage, single linkage, average linkage, and Ward's minimum variance method. **Complete linkage** bases the distance between clusters on the maximum distance between objects in one cluster and another cluster. **Single linkage** bases the distance between clusters on the minimum distance between objects in one cluster and another cluster. **Average linkage** bases the distance between clusters on the mean distance between objects in one cluster and another cluster. **Ward's minimum variance method** bases the distance between clusters on the sum of squares over all variables between objects in one cluster and another cluster.

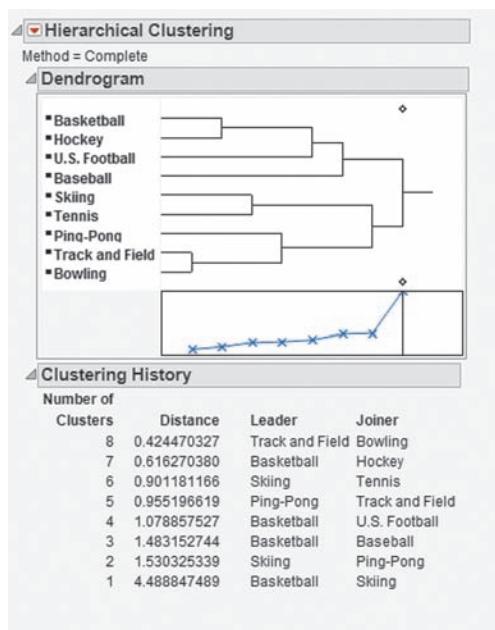
To illustrate cluster analysis, suppose that you wanted to examine the similarities and dissimilarities among various sports. You collect data from a survey on the perceptions of four attributes of nine sports (basketball, skiing, baseball, Ping-Pong, hockey, track and field, bowling, tennis, and U.S. football) and store the data in **Sports**. You define the following seven-point rating scales for the four variables that correspond to the four attributes:

- Movement speed: 1 = fast paced to 7 = slow paced
- Rules: 1 = complicated rules to 7 = simple rules
- Team orientation: 1 = team sport to 7 = individual sport
- Amount of contact: 1 = noncontact to 7 = contact

Figure 17.15 presents the results of a JMP complete linkage cluster analysis based on the mean score of each sport on each rating scale.

FIGURE 17.15

JMP cluster analysis results for the different sports



From examining either the tree diagram (which JMP calls a dendrogram) from left to right or the clustering history, observe that the first two sports that cluster together are track and field and bowling followed by basketball and hockey. Then skiing and tennis join together followed by Ping-Pong merging with track and field and bowling. This process continues until all the sports are merged into one cluster.

When there are three clusters remaining, the sports in the three clusters are {basketball, hockey, U.S. football, and baseball}, {bowling, Ping-Pong, and track and field}, and {tennis and skiing}. The first cluster of {basketball, hockey, U.S. football, and baseball} appears to represent team sports. The second cluster of {bowling, Ping-Pong, and track and field} are slow moving individually contested sports. The third cluster {tennis and skiing} represents fast moving individually contested sports.

Problems for Section 17.5

17.37 Movie companies need to predict the gross receipts of individual movies once the movie has debuted. The following results, stored in **PotterMovies**, are the first weekend gross, the U.S. gross, and the worldwide gross (in \$millions) of the Harry Potter movies.

- Perform a cluster analysis using the complete linkage method on the Harry Potter movies based on the first weekend gross, the U.S. gross, and the worldwide gross (in \$millions).
- What conclusions can you reach about which Harry Potter movies are most similar?

17.38 The file **Cereals** contains the calories, carbohydrates, and sugar, in grams, in one serving of seven breakfast cereals.

- Perform a cluster analysis using the complete linkage method on the cereals based on the calories, carbohydrates, and sugar in grams.
- What conclusions can you reach about which cereals are most similar?

17.39 The file **Protein** contains calorie and cholesterol information for popular protein foods (fresh red meats, poultry, and fish) compiled by the U.S. Department of Agriculture.

- Perform a cluster analysis using the complete linkage method on the protein foods based on the calories and cholesterol, in grams.

- What conclusions can you reach about which protein foods are most similar?
- Perform a cluster analysis using Ward's method on the protein foods based on the calories and cholesterol in grams.
- What conclusions can you reach about which protein foods are most similar?
- Compare the results of (a) and (c). Are there any differences in your conclusions? Explain

17.40 A Pew Research Center survey found that social networking is popular in many nations around the world. The file **GlobalSocialMedia** contains the level of social media networking (measured as the percent of individuals polled who use social networking sites) and the GDP at purchasing power parity (PPP) per capita for each of 25 selected countries. (Data extracted from “Global Digital Communication: Texting, Social Networking Popular Worldwide,” Pew Research Center, bit.ly/sNjsmq.)

- Perform a cluster analysis using the complete linkage method on the nations based on the level of social media networking (measured as the percent of individuals polled who use social networking sites) and the GDP at purchasing power parity (PPP) per capita.
- What conclusions can you reach about which nations are most similar?

17.6 Multidimensional Scaling

Multidimensional scaling (MDS) visualizes objects in a two or more dimensional space, or *map*, with the goal of discovering patterns of similarities or dissimilarities among the objects. One challenge of MDS is to interpret the significance of the dimensions of the map and understand the reasons behind the distance between individual objects or apparent groups of objects. There are two main types of multidimensional scaling: *metric multidimensional scaling* that assumes that the distance between objects is ratio scaled and *nonmetric multidimensional scaling* that assumes that the distance between objects is ordinal scaled.

To perform an MDS analysis, you must determine how to measure the distance between objects (the basis for placing objects in the map) and the number of map dimensions to be interpreted. The most common measure of distance between objects used is the **Euclidean distance** that measures the distance between objects as the square root of the sum of the squared differences between objects over all r dimensions.

EUCLIDEAN DISTANCE (MDS)

$$d_{ij} = \sqrt{\sum_{k=1}^r (X_{ik} - X_{jk})^2} \quad (17.6)$$

where

d_{ij} = distance between object i and object j

X_{ik} = value of object i in dimension k

X_{jk} = value of object j in dimension k

r = number of dimensions

You obtain MDS results in a varying number of dimensions, usually from one dimension to five. The goal of MDS analysis is minimize the number of dimensions used to interpret the results while maximizing the goodness of fit of the results to the original data. The goodness of fit is measured by the **stress statistic** as defined in Equation (17.7).

STRESS STATISTIC

$$\text{Stress} = \sqrt{\frac{\sum_{i,j=1}^m (d_{ij} - \hat{d}_{ij})^2}{\sum_{i,j=1}^m (d_{ij} - \bar{d})^2}} \quad (17.7)$$

where

d_{ij} = distance between objects i and j

\hat{d}_{ij} = the fitted regression value estimated by the multidimensional scaling algorithm from the original data for objects i and j

\bar{d} = mean distance between objects

$m = n(n - 1)/2$

n = number of objects

While the smaller the stress statistic, the better the fit, there is no fixed rule about what constitutes an acceptable value for the stress statistic. Because, as a general rule, the stress statistic decreases as the number of dimensions increases, using many dimensions can cause the stress statistic to approach 0 (a perfect fit), but at the cost of creating a map that could be as complex as the original data itself! Usually, a good rule is to increase dimensions as long as the stress statistic decreases substantially. In many cases, the decrease in the stress statistic begins to level off after the second or third dimension is considered. When this occurs, you can limit yourself to trying to interpret two or three dimensions. Attempting to interpret more than three dimensions in many cases can be extremely challenging.

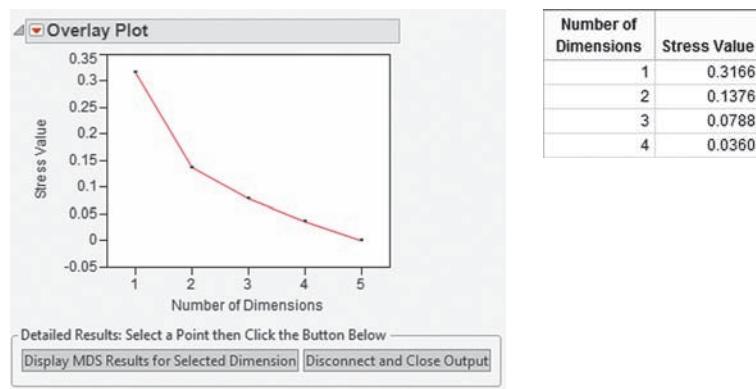
To illustrate MDS analysis, suppose that you wanted to examine the similarities and dissimilarities among various sports. You collect data from a survey on the perceptions of four attributes of nine sports (basketball, skiing, baseball, Ping-Pong, hockey, track and field, bowling, tennis, and U.S. football) and store the data in **Sports**. You define the following seven-point rating scales for the four variables that correspond to the four attributes:

- Movement speed: 1 = fast paced to 7 = slow paced
- Rules: 1 = complicated rules to 7 = simple rules
- Team orientation: 1 = team sport to 7 = individual sport
- Amount of contact: 1 = noncontact to 7 = contact

Figure 17.16 presents the results of the MDS analysis based on the mean score of each attribute for each sport.

FIGURE 17.16

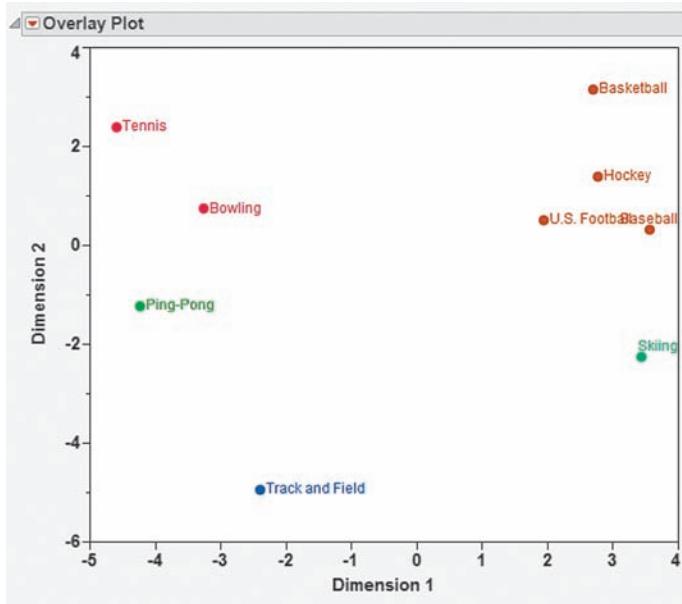
Multidimensional scaling stress results for nine sports



The JMP results reveal a stress statistic of 0.3166 in one dimension, 0.1376 in two dimensions, and 0.0788 in three dimensions. Because there is a large difference in the stress statistic between one and two dimensions but only a small difference in the stress statistic between two and three dimensions, you would choose to begin by interpreting the two-dimensional results shown in Figure 17.17.

FIGURE 17.17

Two-dimensional MDS map for perceptions about nine sports



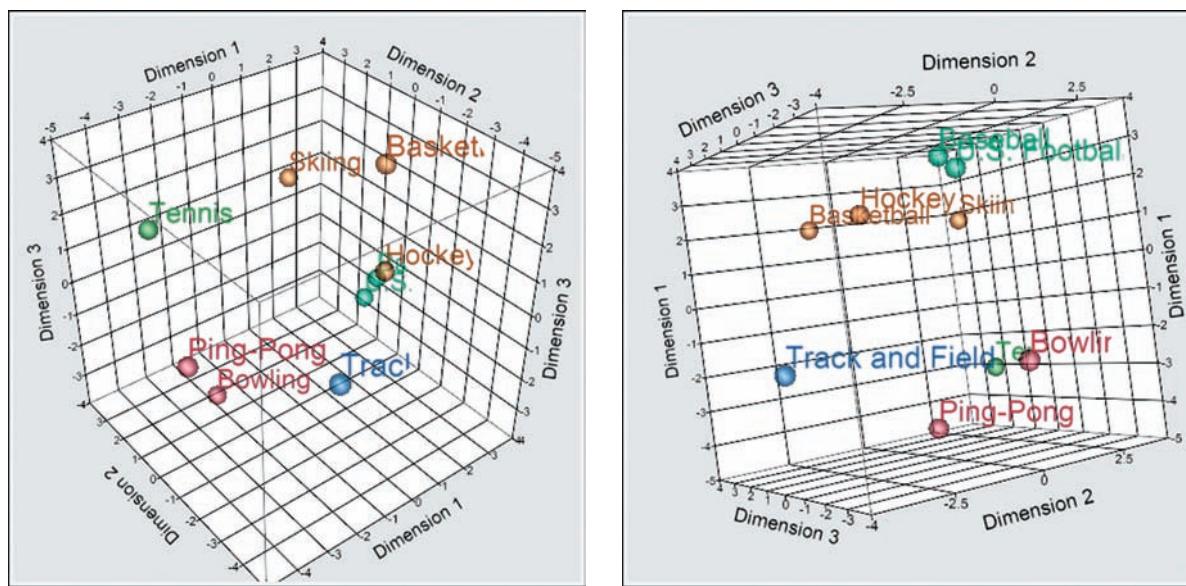
To interpret a two-dimensional map, you look for points that appear close to each other as well as points that appear distant from each other. Although not the case with Figure 17.17, you may need to rotate the map in order to better interpret the dimensions. From Figure 17.17, observe that U.S. football, hockey, and basketball are close to each other. Tennis, Ping-Pong, and bowling are somewhat close to each other, and track and field and skiing are separate from the others.

To best interpret the dimensions separating the sports, observe that if you rotate the map clockwise 45 degrees, one axis appears to separate the team sports (U.S. football, hockey, basketball, and baseball) from the nonteam sports. The other axis appears to separate the fast-paced contact sports (U.S. football, hockey, and basketball) from the slow-paced noncontact sports such as Ping-Pong, bowling, and tennis.

After interpreting the two-dimensional map, you can check to see if interpreting a three-dimensional map yields a better result. Interpreting a three-dimensional map is inherently harder as there are many more ways to examine and rotate the cube-like map. Figure 17.18 shows the original and rotated three-dimensional map. The rotated map seems to show team sports gathering near the “ceiling,” while individual sports gather near the “floor.” As this does not enhance the interpretation, you would use the simpler two-dimensional map for your final analysis.

FIGURE 17.18

Original and rotated three-dimensional MDS maps for perceptions about nine sports



Problems for Section 17.6

- 17.41** Movie companies need to predict the gross receipts of individual movies once the movie has debuted. The following results, stored in **PotterMovies**, are the first weekend gross, the U.S. gross, and the worldwide gross (in \$millions) of the Harry Potter movies.
- Perform a multidimensional scaling analysis on the Harry Potter movies based on the first weekend gross, the U.S. gross, and the worldwide gross (in \$millions).
 - What conclusions can you reach about which Harry Potter movies are most similar?

- 17.42** The file **Cereals** contains the calories, carbohydrates, and sugar, in grams, in one serving of seven breakfast cereals.
- Perform a multidimensional scaling analysis on the cereals based on the calories, carbohydrates, and sugar in grams.
 - What conclusions can you reach about which cereals are most similar?

- 17.43** The file **Protein** contains calorie and cholesterol information for popular protein foods (fresh red meats, poultry, and fish) compiled by the U.S. Department of Agriculture.

- Perform a multidimensional scaling analysis on the protein foods based on the calories and cholesterol, in grams.
- What conclusions can you reach about which protein foods are most similar?

- 17.44** A Pew Research Center survey found that social networking is popular in many nations around the world. The file **GlobalSocialMedia** contains the level of social media networking (measured as the percent of individuals polled who use social networking sites) and the GDP at purchasing power parity (PPP) per capita for each of 21 selected countries. (Data extracted from “Global Digital Communication: Texting, Social Networking Popular Worldwide,” Pew Research Center, bit.ly/sNjsmq.)

- Perform a multidimensional scaling analysis on the nations based on the level of social media networking (measured as the percent of individuals polled who use social networking sites) and the GDP at purchasing power parity (PPP) per capita.
- What conclusions can you reach about which nations are most similar?

USING STATISTICS

Finding the Right Lines at WaldoLands, Revisited

Business analytics has various uses at WaldoLands. Descriptive analytics help report the current status of the theme park and handle the logistics of an enterprise as complex as a small city. Predictive analytics can be used to anticipate demands for attractions, food, or other services, as well as discover patterns about park patrons that could be used to enhance advertising efforts or generate additional revenues. Working with “big data” that tracks spending patterns and links theme

park data to credit file information, other predictive methods could identify which groups of hotel guests are the theme park’s “best” customers and might be candidates for future targeted marketing efforts. Prescriptive analytics, not discussed in this chapter, could be used to help plan the strategic investments in the park and to anticipate the effects of a changing economic climate or new competition.



REFERENCES

1. Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. London: Chapman and Hall, 1984.
2. Cox, T. F., and M. A. Cox. *Multidimensional Scaling*, Second ed. Boca Raton, FL: CRC Press, 2010.
3. Everitt, B. S., S. Landau, and M. Leese. *Cluster Analysis*, Fifth ed. New York: John Wiley, 2011.
4. Few, S. *Information Dashboard Design: Displaying Data for At-a-Glance Monitoring*, Second ed. Burlingame, CA: Analytics Press, 2013.
5. Hakimpoor, H., K. Arshad, H. Tat, N. Khani, and M. Rahmandoust. "Artificial Neural Network Application in Management." *World Applied Sciences Journal*, 2011, 14(7): 1008–1019.
6. *JMP Version 10*. Cary, NC: SAS Institute. 2012.
7. Lindoff, G., and M. Berry. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Hoboken, NJ: Wiley Publishing, Inc., 2011.
8. Loh, W. Y. "Fifty Years of Classification and Regression Trees." *International Statistical Review*, 2013.
9. *Microsoft Excel 2013*. Redmond, WA: Microsoft Corp., 2012.
10. *Tableau Public Version 8*. Seattle, WA: Tableau Software, 2013.
11. Tufte, E. *Beautiful Evidence*. Cheshire, CT: Graphics Press, 2006.
12. Turban, E., R. Sharda, J. Aronson, and D. King. "Networks for Data Mining, Online Chapter 6," *Business Intelligence: A Managerial Approach*. Upper Saddle River, NJ: Prentice Hall, 2008.

KEY EQUATIONS

Akaike Information Criterion (AIC)

$$\text{AIC} = 2k - 2 \ln(L) \quad (17.1)$$

Akaike Information Criterion corrected (AIC_c)

$$\text{AIC}_c = \text{AIC} + \frac{2k(k + 1)}{n - k - 1} \quad (17.2)$$

LogWorth

$$\text{LogWorth} = -\log_{10}(p\text{-value}) \quad (17.3)$$

Hyperbolic Tangent Function

$$\frac{e^{2x} - 1}{e^{2x} + 1} \quad (17.4)$$

Euclidean Distance (Cluster Analysis)

$$d_{ij} = \sqrt{\sum_{k=1}^r (X_{ik} - X_{jk})^2} \quad (17.5)$$

Euclidean Distance (MDS)

$$d_{ij} = \sqrt{\sum_{k=1}^r (X_{ik} - X_{jk})^2} \quad (17.6)$$

Stress Statistic

$$\text{Stress} = \sqrt{\frac{\sum_{i, j=1}^m (d_{ij} - \hat{d}_{ij})^2}{\sum_{i, j=1}^m (d_{ij} - \bar{d})^2}} \quad (17.7)$$

KEY TERMS

Akaike information criterion (AIC) 692
 average linkage 692
 bullet graph 677
 business analytics 675
 classification and regression trees 683
 cluster analysis 691
 complete linkage 692
 dashboard 676
 data mining 682
 data discovery 678
 descriptive analytics 675

drill-down 678
 Euclidean distance 692
 gauges 677
 Gini impurity 683
 hierarchical clustering 691
 hyperbolic tangent function 688
 k-means clustering 691
 multidimensional scaling (MDS) 693
 multilayer perceptrons (MLPs) 688
 neutral networks 688
 predictive analytics 675

prescriptive analytics 675
 processing elements 688
 single linkage 692
 slicers 679
 sparklines 676
 stress statistic 694
 training data 689
 treemap 677
 Ward's minimum variance method 692

CHECKING YOUR UNDERSTANDING

17.45 What is the difference between a gauge and a bullet graph, and what are the advantages and disadvantages of each?

17.46 How are sparklines different from time-series plots?

17.47 What is the purpose of a treemap?

17.48 How do classification trees differ from regression trees?

17.49 How do classification and regression tree models differ from neural net models?

17.50 How does cluster analysis differ from multidimensional scaling?

CHAPTER REVIEW PROBLEMS

17.51 The file **DomesticBeer2** contains the number of calories per 12 ounces and number of carbohydrates (in grams) per 12 ounces for a sample of 15 of the best-selling domestic beers in the United States. (Data extracted from www.beer100.com/beer-calories.htm.)

- Visually evaluate the number of calories per 12 ounces for each beer by constructing a bullet graph. For the purposes of comparison, consider calories below 100 as low, between 100 and 160 as medium, and above 160 as high.
- Visually evaluate the number of carbohydrates (in grams) per 12 ounces for each beer by constructing a bullet graph. For the purposes of comparison, consider carbohydrates below 10 grams as low, between 10 and 14 grams as medium, and above 14 grams as high.
- What preliminary conclusions can you reach about the number of calories and carbohydrates in the beers?
- Why would constructing sets of gauges for the calories and carbohydrates be a less effective means of visualizing these data?

17.52 The file **Currency2** contains the value of the Canadian dollar, English pound, and Euro for one U.S. dollar from 2002 to 2012.

- Construct sparklines for the value of the U.S. dollar in terms of the Canadian dollar, English pound, and Euro.
- What conclusions can you reach about the value of the U.S. dollar in terms of the Canadian dollar, English pound, and Euro from 2002 to 2012.

17.53 The production of wine is a multibillion-dollar worldwide industry. In an attempt to develop a model of wine quality as judged by wine experts, data were collected from red and white wine variants of Portuguese “Vinho Verde” wine. (Data extracted from P. Cortez, Cerdeira, A., Almeida, F., Matos, T., and Reis, J., “Modeling Wine Preferences by Data Mining from Physiochemical Properties,” *Decision Support Systems*, 47, 2009, pp. 547–553 and bit.ly/9xKIEa.) The population of 6,497 wines is stored in **VinhoVerde Population**.

- Using half the data as the training sample and the other half of the data as the validation sample, develop a classification tree model to predict the probability that the wine is red. (Consider the entire set of variables in your analysis.)
- What conclusions can you reach about the probability that the wine is red.
- Repeat (a) using a neural net model.
- Compare the results of (a) and (c).

17.54 Referring to the data in Problem 17.53,

- Using half the data as the training sample and the other half of the data as the validation sample, develop a regression tree model to predict wine quality. (Consider the entire set of variables in your analysis.)
- What conclusions can you reach about wine quality?
- Repeat (a) using a neural net model.
- Compare the results of (a) and (c).

17.55 The file **FTMBA** contains a sample of 2012 top-ranked full-time MBA programs. Variables included are mean starting salary upon graduation (\$), percentage of students with job offers within three months of graduation, program cost (\$), and total number of students per program. (Data extracted from buswk.co/25d1ZC.)

- Using all the data as the training sample, develop a regression tree model to predict the mean starting salary upon graduation
- What conclusions can you reach about the mean starting salary upon graduation?
- Using half the data as the training sample and the other half of the data as the validation sample, develop a regression tree model to predict the mean starting salary upon graduation.
- What differences exist in the results of (a) and (c)?
- Repeat (c) using a neural net model.
- Compare the results of (c) and (e).

17.56 A specialist in baseball analytics is interested in determining which variables are important in predicting a team’s wins in a given baseball season. He has collected data in **BB2012** that includes the number of wins, ERA, saves, runs scored, hits allowed, walks allowed, and errors for the 2012 season.

- Using half the data as the training sample and the other half of the data as the validation sample, develop a regression tree model to predict the number of wins.
- What conclusions can you reach about the number of wins?
- Repeat (a) using a neural net model.
- Compare the results of (a) and (c).

17.57 Nassau County is located approximately 25 miles east of New York City. Data in **GlenCove** are from a sample of 30 single-family homes located in Glen Cove. Variables included are the fair market value, land area of the property (acres), interior size of the house (square feet), age (years), number of rooms, number of bathrooms, and number of cars that can be parked in the garage.

- Using all the data as the training sample, develop a regression tree model to predict the fair market value.

- b. What conclusions can you reach about the fair market value?
- c. Using half the data as the training sample and the other half of the data as the validation sample, develop a regression tree model to predict the fair market value.
- d. What differences exist in the results of (a) and (c)?
- e. Repeat (c) using a neural net model.
- f. Compare the results of (c) and (e).

17.58 A market research study has been conducted by a travel website that specializes in restaurants with the business objective of determining which types of foods are perceived to be similar and which are perceived to be different. The following 10 types of foods were studied:

Japanese	Mandarin (Chinese)
Cantonese (Chinese)	American
Szechuan (Chinese)	Spanish
French	Italian
Mexican	Greek

The mean values of each food on the scales of

- Bland (1) Spicy (7)
- Light (1) Heavy (7)
- Low calories (1) High calories (7)

are stored in **Foods**.

- a. Perform a cluster analysis on the types of foods.
- b. Perform a multidimensional scaling analysis on the types of foods.
- d. What conclusions can you reach about which types of foods are most similar?

17.59 A specialist in baseball analytics seeks to study which baseball teams were most similar in 2012. He has collected data in **BB2012** related to ERA, saves, runs scored, hits allowed, walks allowed, and errors for the 2012 season.

- a. Perform a cluster analysis on the baseball teams.
- b. Perform a multidimensional scaling analysis on the baseball teams.
- c. What conclusions can you reach about which baseball teams were similar in 2012?

CASE FOR CHAPTER 17

The Mountain States Potato Company

On page 623, you studied the Mountain States Potato Company which needed to determine why the percentage of solids in the filter cake that it sells was below its historical

value. Construct a regression tree model for the percentage of solids in the filter cake and include the results in the report that is to be submitted to the president of the company.

CHAPTER 17 SOFTWARE GUIDE

INTRODUCTION

Chapter 17 discusses a number of statistical methods not included in or weakly supported by Microsoft Excel and Minitab and uses JMP, a separate statistical application, for its predictive analytics examples. For these reasons, this *Software Guide* presents instructions for using JMP as well as Tableau Public 8, a browser-based (“cloud-based”) application that specializes in interactive data visualization, in addition to Microsoft Excel. (There are no Minitab instructions for Chapter 17 methods, although some of the Chapter 2 and Chapter 3 Minitab instructions can be adapted for descriptive analytics.) Use Table SG17.1 as a guide for choosing which program to use for which method. As explained in the SHORT TAKES for Chapter 17, you can download a free 30-day, full-featured trial version of JMP from the SAS Institute. You can also download for free upon registration Tableau Public 8, a “cloud-based” Internet application, from Tableau Software and the Gauge and Treemap Apps for Office for Excel 2013 or Office 365 for Microsoft Windows from the Microsoft Office Store.

TABLE SG17.1

Software Guide Instructions for Chapter 17 Methods

Method	Excel	Tableau	JMP
Sparklines	(1)		
Gauges	Gauge App		
Bullet Graph	•	•	
Treemaps	Treemap App	•	•
Drill-down	•		•
Slicers	(2)		•
Classification and Regression Trees			•
Neural Networks			•
Cluster Analysis			•
Multidimensional Scaling		(3)	

Notes:

- (1) Also available in Excel 2010.
- (2) Not available in OS X Excel 2011 or Office 365 for Mac.
- (3) Requires additional downloads of a JMP add-in (free upon registration) and R, the free, open-source statistical package. (See the SHORT TAKES for Chapter 17 for further details.)

This book does not include a complete orientation to either JMP or Tableau Public 8 (as was done for Excel and Minitab in the opening chapters). If you are a computing novice you may find using JMP or Tableau Public initially challenging. While both programs include extensive tutorials, mastery of basic computing skills is assumed and, in the case of JMP, the tutorials cover basic statistical methods and not the advanced analytics methods used in this chapter.

SG17.1 DESCRIPTIVE ANALYTICS

Sparklines

In-Depth Excel Use Sparklines.

For example, to create the Figure 17.2 sparklines display, open to the **DATA worksheet** of the **WL_WaitHistory workbook**. In this worksheet, ride names are in column A and the historical wait times data by half-hours are in Columns C through W. Select cell range **C3:W16** and:

1. Select **Insert → Line** (in the **Sparklines** group).
2. In the Create Sparklines dialog box, enter **B3:B16** as the **Location Range** and click **OK**.
3. Select cell **B3** (or any other column B cell that contains a sparkline). Then select **Design** (in the Sparkline Tools) → **Axis**.
4. In the Axis drop-down gallery, click **Same for all Sparklines** in the **Vertical Axis Minimum Value Options**. Click **Axis** a second time and click **Same for all Sparklines** in the **Vertical Axis Maximum Value Options**.
5. Check **Last Point** (in the **Show** group of the Design tab).

In the Data worksheet, the row heights of rows 3 through 16 have already been increased to 20 units. You can increase the row height further or widen column B by selecting the rows (or column), right-clicking and selecting **Row Height** (or **Column Width**) from the shortcut menu, entering a new value, and clicking **OK**.

Gauges

In-Depth Excel Use the **Gauge App** (requires being signed in to the Microsoft Office Store).

For example, to construct a dashboard of six gauges, equivalent to the one shown in Figure 17.3 on page 677, open to the **TopSixDATA worksheet** of the **WL_WaitData workbook** and:

1. Select **Insert → Apps for Office** and click **Gauge** in the Apps for Office gallery. If selecting Gauge from the Apps for Office dialog box (and not from the Recently Used Apps list), also click **Insert**.
2. When the RE:VISION Gauge panel appears, click **Insert Gauges**.
3. In the Select Data dialog box, enter **C2:C7** and click **OK**.
4. Drag and resize the panel of six gauges until they can be easily read.

By default, gauges constructed by the Gauge App have a minimum value of 0 and a maximum of 100 and include three gauge zones: green, defined as values between 0 and 25; yellow, defined as values between 40 and 60, and red, defined as values between 75 and 100. These default values create a gauge that effectively has five zones: green, white (values from 25 to 40), yellow, white again (values from 60 to 75), and red. By adjusting the scope of the yellow range, you can create gauges that contain three or four zones. The Figure 17.3 set of gauges contains five zones that differ in scope from the default ranges.

To adjust the zones, click the **circular gear icon** at the top of panel. In the Settings panel:

5. Enter **45** as the Yellow Range **Min Value**.
6. Enter **65** as the Yellow Range **Max Value**.
7. Enter **85** as the Red Range **Min Value**.
8. Click the **back icon** (left-facing arrow icon) at the top of the Settings panel.

If you use an Excel version older than Excel 2011 (or a newer version not signed into the Microsoft Office Store), open to the **Gauge worksheet** of the **GaugeBullet workbook** to view a non-modifiable version of the set of gauges that steps 1 through 8 construct. Also note that various third-party vendors offer Excel add-ins that can add gauges to current and older versions of Excel. For example, the Figure 17.3 set of gauges are adapted and enhanced versions of gauges that the free version of the BeGraphic add-in (**begraphic.com**) can create.

Gauges can also be simulated in any Excel version by overlaying a pie chart on a donut chart (a chart type not discussed in this book). In the pie chart, all slices of the pie chart are set to invisible except for the slice that represents the needle, and the background of the pie chart is also set to be invisible. In the donut chart, the lower half of the donut is made invisible to form an 180-degree arc that serves as the legend for the zones of the gauge. The pie is then moved over and scaled to the donut chart to complete the gauge. Open to the **COMPUTE worksheet** of the **RadialGauge workbook** to see an illustration of this method and to experiment with the user-changeable values in the tinted cells in columns B and C.

Bullet Graph

In-Depth Excel Use the **BulletGraph worksheet** of the **GaugeBullet workbook** as a model for simulating a bullet graph.

To construct a simulated bullet graph in Excel, you create a bar chart of the variable being graphed with a transparent background and overlay this chart on a bar chart that displays the colored zones. For example, to construct a chart similar to the bullet graph shown in Figure 17.3 on page 677, open to the **waitDATA worksheet** of the **WL_WaitData workbook** and:

1. Select cell range **B1:C15**.
2. Select **Insert**, then the **bar chart icon**, and select the **first 2-D Bar** galley item (**Clustered Bar**).
3. In the newly constructed bar chart, turn off the gridlines (and, in Excel 2010, legend), by using the instructions in Appendix Section B.6.
4. Right-click in the whitespace to the right of the chart title (upper right corner of chart in Excel 2010) and click **Format Chart Area** in the shortcut menu.
5. In the Fill part of the Format Chart Area pane (dialog box in Excel 2010), click **No fill**. (In Excel 2010, also click **Close**.) The background of the chart becomes transparent. In Excel 2010, also right-click the plot area, click **Format Plot Area**, and in the Fill part of the Format Plot Area dialog box, click **No fill** and then click **Close**.

Next, construct the bar chart that will serve as the colored zones for the bullet graph.

6. In the cell range **D2:D6**, enter the values **25, 20, 20, 20**, and **15**, to define the five zones of the Figure 17.3 bullet graph. Then select this edited cell range **D2:D6**.

7. Select **Insert**, then the **bar chart icon**, and select the **second 2-D Bar** galley item (**Stacked Bar**).

8. In the newly constructed bar chart, turn off the gridlines (and, in Excel 2010, legend) by using the instructions in Appendix Section B.6.

9. Right-click in the whitespace to the right of the chart title (upper right corner of chart in Excel 2010) and click **Select Data** in the shortcut menu.

10. In the Select Data Source dialog box, click **Switch Row/Column** and then click **OK**. A chart of five simple bars becomes a chart of one stacked bar with five parts.

11. Right-click the one stacked bar and click **Format Data Series** in the shortcut menu. In Excel 2013, click the icon that looks like a small histogram; otherwise click **Series Options** in the left pane. In the Series Options part of the Format Data Series pane (dialog box in Excel 2010), change **Gap Width** to **0%**. (In Excel 2010, also click **Close**.)

12. Change the coloring of the stacked bars. In Excel 2013, select **Design → Change Colors** and in the gallery click one of the color spectrums. In Excel 2010, select **Design** and click one of the items in the **Chart Styles** group. Be sure to choose a set of colors that does not include the color used for the bars in the bar chart you constructed using steps 1 through 5.

13. Right-click the horizontal chart axis and click **Format Axis** in the shortcut menu.

14. In the Axis Options of the Format Axis pane (dialog box in Excel 2010), enter **100** as the **Maximum**. In Excel 2010, first click **Fixed** in the Maximum line first, then enter **100**, and then click **Close**.

15. Adjust the size of the chart, as necessary, by clicking a corner of the bar chart frame and then dragging that corner to resize the chart.

16. Right-click the chart border and select **Send to Back → Send to Back** in the shortcut menu.

17. Drag the bar chart with the transparent background over the stacked bar chart and adjust so that the zeroes on the horizontal axis of both charts coincide. Then adjust the width of that bar chart so that all other horizontal axis numbers that the two charts share coincide.

For other problems, you need to identify the maximum value and enter the proper set of values in new column in order to correct the correct stacked bar chart that serves to display the zones for the bullet graph.

Tableau Public Modify the bar chart created automatically from the data.

For example, to construct the Figure 17.3 bullet graph of wait times, select **File → New** and:

1. Select **Data → Connect to Data**. In the Connect to Data panel, click **Microsoft Excel**.
2. In the Open dialog box, navigate to the location of the **WL_WaitData Excel workbook**, select that file, and then click **Open**.
3. In the Excel Workbook Connection dialog box, click **waitDATA** in the Step 2 list box and click **OK**.

Tableau Public displays a Data pane similar to one shown below next to an empty worksheet.



From the Data pane:

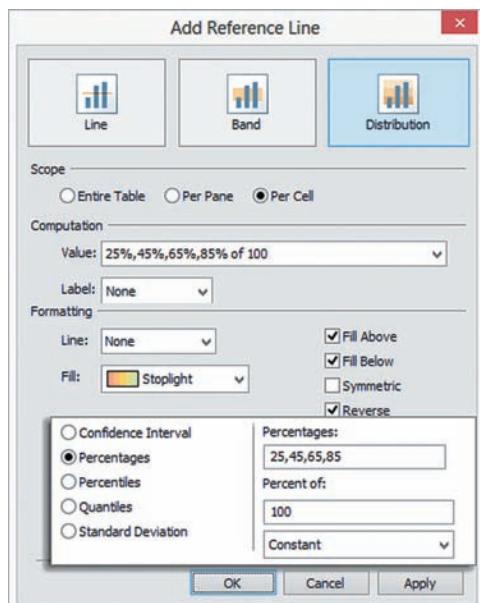
4. Drag **Ride** in the Dimensions group and drop it in the **Rows** box of the new worksheet area.
5. Drag **Wait** in the Measures group and drop it in the **Columns** box. (**Wait** changes to **SUM(Wait)**.)

Tableau constructs a simple bar chart of wait times. To change this chart into a bullet graph:

6. Right-click the Wait axis, and in the shortcut menu click **Add Reference Line**.

In the Add Reference Line dialog box (shown below):

7. Click **Distribution**.
8. Click **Per Cell** as the Scope.
9. Click the **Value** entry to reveal the Value entries (shown as inset below).
10. In the Value entries, click **Percentages**, enter **25,45,65,85** as the **Percentages**, select **Constant** from the drop-down list, and enter **100** as the **Percent of** value. (Although the drop-down list appears *after* the **Percent of** entry, select Constant before entering the value 100.)
11. In the **Label** drop-down list, select **None**.
12. In the **Fill** drop-down list, select **Stoplight**.
13. Check **Fill Above**, **Fill Below**, and **Reverse**.
14. Click **OK**.



You can adjust the formatting of axis labels or the chart title by right-clicking over the element and clicking **Format** in the shortcut menu and making adjustments in the Format pane. The Format pane appears over the same space used to display the Data pane. To go back to the Data pane, close the Format pane.

Treemap

In-Depth Excel Use the **Treemap app** (requires being signed in to the Microsoft Office Store).

For example, to construct a treemap of WaldoLands social media comments grouped by “Land,” shown at left in Figure 17.4, on page 677, open to the **DATA worksheet** of the **WL_SocialData workbook** and:

1. Select **Insert → Apps for Office** and click **Treemap** in the Apps for Office gallery. If selecting Treemap from the Apps for Office dialog box (and not from the Recently Used Apps list), also click **Insert**.

In the Treemap panel:

2. Click **Name list** and in the Select Data dialog box enter **A2:B15** and click **OK**. A treemap begins to take shape in the Treemap panel.
3. Click **Size** and in the Select Data dialog box enter **C2:C15** and click **OK**.
4. Click **Color** (under **Size**), and in the Select Data dialog box enter **D2:D15** and click **OK**.

If the treemap displayed does not use the red-to-blue spectrum, click the **spectrum icon** (under Title) and click the red-to-blue (third) spectrum.

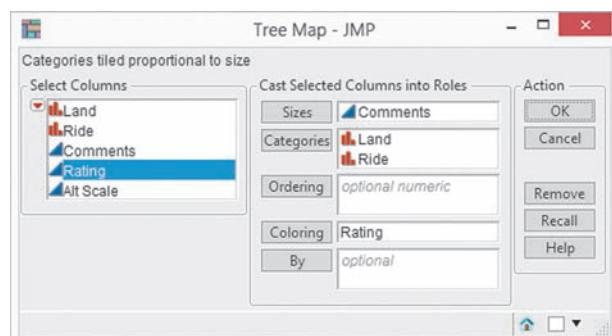
To construct the treemap shown at right in Figure 17.4, repeat steps 1 through 4, entering **B2:B7** in step 2, **C2:C7** in step 3, and **D2:D7** in step 4. If you use an Excel version older than Excel 2011 (or a newer version not signed into the Microsoft Office Store), open to the **Nested** and **Simple** worksheets of the **Treemap workbook** to view nonmodifiable versions of the Figure 17.4 treemaps.

JMP Use Treemap.

For example, to construct a treemap of WaldoLands social media comments grouped by “Land,” similar to the one shown at left in Figure 17.4 on page 677, open the **WL_SocialData Excel workbook**. Select **File → Open**. In the Open Data File dialog box, navigate to the location of the file, select the file, and then click **Open**. In the DATA - JMP window:

1. Select **Graph → Tree Map**.

In the Tree Map - JMP dialog box (shown below):



2. Drag **Land** to the **Categories** box.
3. Drag **Ride** to the **Categories** box.
4. Drag **Comments** to the **Sizes** box.
5. Drag **Rating** to the **Coloring** box.
6. Click **OK**.

Adjust the size of the treemap as necessary to clearly display the labels. By default, JMP colors the treemap using blue for the unfavorable ratings and red for the favorable ratings, the inverse of the colorings in Figure 17.4. To change the color spectrum, click the drop-down button that is part of the chart title, click **Color Theme**, and then click one of the submenu choices. Note that there is no predefined spectrum that uses red for the lowest values.

To construct a treemap similar to the one shown at right in Figure 17.4, in the TopSix DATA - JMP window, repeat steps 1 through 6, skipping step 2.

Tableau Public

Use the treemap feature. For example, to construct a treemap of WaldoLands social media comments grouped by “Land,” similar to the one shown at left in Figure 17.4 on page 677, select **File → New** and:

1. Select **Data → Connect to Data**. In the Connect to Data panel, click **Microsoft Excel**.
2. In the Open dialog box, navigate to the location of the **WL_SocialData Excel workbook**, select that file, and then click **Open**.
3. In the Excel Workbook Connection dialog box, click **DATA** in the Step 2 list box and click **OK**.

Tableau Public displays a Data pane next to an empty worksheet. From the Data pane:

4. Drag **Comments** in the Measures group and drop it on the **Size icon** in the Marks area.
5. Drag **Rating** in the Measures group and drop it on the **Color icon** in the Marks area.
6. Drag **Land** in the Dimensions group and drop it on the **Label icon** in the Marks area of the new worksheet area.
7. Drag **Ride** in the Dimensions group and drop it on the **Label icon** in the Marks area.

Tableau Public constructs a treemap and updates the Marks area (called the “Marks card” by Tableau) and adds a SUM(rating) area similar to these areas:



To change the red-to-green spectrum to colors less likely to be confused:

8. Double-Click the SUM(Rating) color spectrum.

In the Edit Colors [(Rating)] dialog box:

9. Select **Red-Blue Diverging** from the **Palette** drop-down list.

10. Click **OK**.

Change the dimensions of the treemap to allow all labels to be displayed. To change a dimension, move the mouse pointer over an edge and then drag the edge to adjust. To adjust the formatting of the labels, select **Format → Font**, and in the Font pane change the font attributes for **Pane** in the default group.

Tableau Public allows the interactive collapse of a level of data. For example, to collapse the rides into their lands, right-click **Ride** in the Marks area, and in the shortcut menu click **Attribute**. The treemap changes to a three area map, one for each of WaldoLands’ three lands. To restore the original map, right-click **Ride** in the Marks area, and in the shortcut menu click **Dimension**.

There are two ways to construct the treemap shown at right in Figure 17.4. You can repeat steps 1 through 10, clicking **Top6DATA** in step 3 and skipping step 4, or you can use the filter function to alter the treemap that is produced by the original steps 1 through 10. To use the filter function, first follow steps 1 through 10 above. Right-click **Ride** in the Marks area, and in the shortcut menu click **Filter**. In the General tab of the Filter [Ride] dialog box, clear the check boxes for the rides that are not part of the top six attractions and then click **OK**.

Data Discovery

In-Depth Excel

Use **PivotTables** and **Slicer**. For drill-down, construct a PivotTable, adapting the Section EG2.6 “Multidimensional Contingency Tables” as necessary. (While those instructions discuss using only categorical variables, the same set of instructions can be used for a mix of categorical and numerical variables.) Then click on the “+” buttons that precede the row categories to expand the table one-level deeper. Figure 17.5 on page 678 illustrates this operation.

To reveal the data for variables not initially included in the initial PivotTable, as illustrated by Figure 17.6 on page 679, include the cell range of those variables when first defining the PivotTable and then later click on a cell that contains the value of interest as also discussed on page 95.

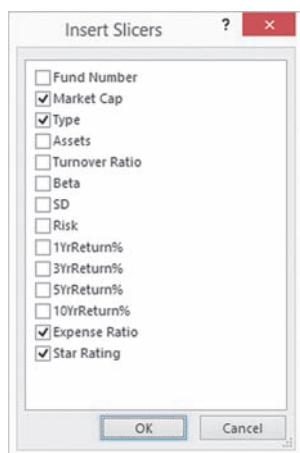
To construct the Figure 17.6 slicer dashboard, first construct a PivotTable using the *In-Depth Excel* “The Contingency Table” instructions on page 86. Click cell **A3** in the PivotTable and:

1. Select **Insert → Slicer**.

In the Insert Slicers dialog box (shown on page 704):

2. Check **Market Cap**, **Type**, **Expense Ratio**, and **Star Rating**.

3. Click **OK**.



4. In the worksheet, drag the slicers to reposition them. If necessary, resize slicer panels as you would resize a window.

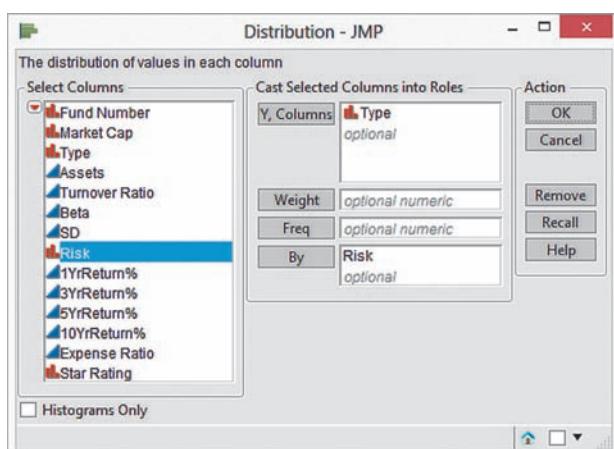
Click the value buttons in the slicers to explore the data. For example, to create the display shown at left in Figure 17.9 on page 680 that answers the question about the attributes of the fund with the highest expense ratio, click **6.97** in the **Expense Ratio** slicer.

When you click a value button, the icon at the top right of the slicer changes to include a red X (as can be seen in both Expense Ratio slicers in Figure 17.9). Click this icon to reset the slicer. When you click a value button, value buttons in *other* slicers may become dimmed (as have buttons such as Value, Mid-Cap, Small, Five, Four, One, and Two in Figure 17.9). Dimmed value buttons represent values that are not found in the currently “sliced” data, and if you click a dimmed value button, the PivotTable will be empty and show no values.

JMP Use Graph Builder.

Summary charts can be used as the basis for both drill-down and slicer-like operations in JMP. For example, to create a drill-down similar to the example that Figure 17.6 illustrates, open the **Retirement Funds Excel workbook**. Select **File → Open**. In the Open Data File dialog box, navigate to the location of the file, select the file, and then click **Open**. In the DATA - JMP window, select **Analyze → Distribution** and in the Distribution - JMP dialog box (shown below):

1. Drag **Type** to the **Y, Columns** box.
2. Drag **Risk** to the **By** box.
3. Click **OK**.



JMP constructs a panel of three histograms that show the distribution of growth and value funds for each type of risk (low, average,

and high). Double-click a histogram bar to display all of the variables for funds that have the bar’s combination of type and risk values.

To create a slicer-like display, in the DATA - JMP window, again select **Analyze → Distribution** and in the Distribution - JMP dialog box:

1. Drag **Type** to the **Y, Columns** box.
2. Drag **Market Cap** to the **Y, Columns** box.
3. Drag **Star Rating** to the **Y, Columns** box
4. Drag **Expense Ratio** to the **Y, Columns** box
5. Click **OK**.

JMP constructs a panel that contains four histograms, one for each variable. (The display for the numerical variable **Expense Ratio** also includes a box-and-whisker plot.) Click a specific bar to display the proportion of the bars in the other histograms related to that bar’s value. For example, when you click the **Value** bar in the **Type** histogram, you can see that among value funds large is the most frequent market cap and that three stars is the most frequent star rating.

SG17.2 PREDICTIVE ANALYTICS

There are no software guide instructions for this section.

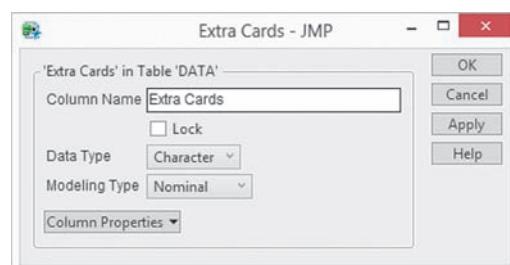
SG17.3 CLASSIFICATION and REGRESSION TREES

Classification Tree

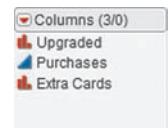
JMP Use Partition

For example, to perform the Figure 17.10 classification tree analysis for predicting the proportion of credit card holders who would upgrade to a premium card, open the **CardStudy Excel workbook**. Select **File → Open**. In the Open Data File dialog box, navigate to the location of the file, select the file, and then click **Open**. Because the Upgraded and Extra Cards variables have been coded with values 0 and 1, JMP mistakes these categorical variables as numerical variables (and would perform an incorrect analysis on the data).

To change the variable type of Extra Cards to categorical, right-click the Extra Cards column and click **Column Info** in the shortcut menu. In the Extra Cards - JMP dialog box (shown below), select **Character** from the **Data Type** drop-down list and click **OK**. (The Modeling Type changes from Continuous to Nominal.) To change Upgraded to a categorical variable, right-click the Upgraded column, click **Column Info**, and select **Character** as the **Data Type** and click **OK**.



Verify these operations by examining the icons that appear in the Columns panel to the left of the JMP worksheet. If these icons, which JMP calls *modeling type icons*, appear as shown at right, the Upgraded and Extra Cards variables have been properly identified as categorical variables (with a nominal scale).



While at this point, you could begin your analysis, the results will use the values 0 and 1, whose meanings may be misinterpreted, for the two categorical variables. Better would be results that use the easily understood values No and Yes. To recode Extra Cards in this way, select the Extra Cards column and:

1. Select Cols → Recode.
2. In the Recode - JMP dialog box (shown below), enter **No** as the New Value for 0, enter **Yes** as the New Value for 1, and click **OK**.



To recode Upgraded, select the Upgraded column and repeat the previous steps 1 and 2.

With the data properly prepared, in the Data - JMP window, select **Analyze → Modeling → Partition**. In the Partition dialog box (shown below):

3. Drag **Upgraded** to the **Y, Response** box.
4. Drag **Purchases** to the **X, Factor** box.
5. Drag **Extra Cards** to the **X, Factor** box.
6. Click **OK**.



In the new DATA - Partition of Upgraded - JMP dialog box:

7. Click the drop-down button to the left of the diagram title and click **Split Best**. Repeat this step until clicking Split Best no longer has any effect on the tree diagram.
8. If the contents of the diagram do not match Figure 17.10, click the drop-down button to the left and then select **Display Options**. To match Figure 17.10, all choices on the Display Options submenu should be checked, except the last two choices, **Show Split Candidates** and **Sort Split Candidates**. If necessary, click unchecked choices, one at a time, until all but the last choice are checked.
9. Click **Color Points**. The Color Points button disappears and the “No” points in the plot appear in red and the “Yes” points appear in blue.

At any point, click **Prune** to remove the last split operation. To enhance the display of the points in the plot, right-click a point, then click **Marker Size** from the shortcut menu and click one of the size choices.

Regression Tree

JMP Use Partition.

For example, to perform the Figure 17.11 regression tree analysis for predicting the sales of OmniPower bars, open the **OmniPower Excel workbook**. Select **File → Open**. In the Open Data File dialog box, navigate to the location of the file, select the file, and then click **Open**. Because JMP properly identifies Sales, Price, and Promotion as numerical variables, there is no need to change variable types as is done in the preceding classification tree example. In the DATA - JMP window, select **Analyze → Modeling → Partition**. In the Partition dialog box:

1. Drag **Sales** to the **Y, Response** box.
2. Drag **Price** to the **X, Factor** box.
3. Drag **Promotion** to the **X, Factor** box.
4. Click **OK**.

In the new DATA - Partition of Upgraded - JMP dialog box:

5. Click the drop-down button to the left of the title **Partition for Sales** and click **Split Best**. Repeat this step until clicking Split Best no longer has any effect on the tree diagram.

At any point, click **Prune** to remove the last split operation. To enhance the display of the points in the plot, right-click a point, then click **Marker Size** from the shortcut menu and click one of the size choices.

SG17.4 NEURAL NETWORKS

JMP Use Neural.

For example, to perform the Figure 17.13 MLP analysis for classifying credit card holders who would be likely to upgrade to a premium card, open the **CardStudy Excel workbook**. Select **File → Open**. In the Open Data File dialog box, navigate to the location of the file, select the file, and then click **Open**. As discussed in Section SG17.3, because the Upgraded and Extra Cards variables have been coded with values 0 and 1, JMP mistakes these categorical variables for numerical variables (and would perform an incorrect analysis on the data).

First use the instructions in Section SG17.3 to change the data type (to Character) for these two variables. Then use the instructions in that section for recoding the values of the two variables as No and Yes. With the data properly prepared, select **Analyze → Modeling → Neural**. In the Neural JMP dialog box (similar to the Partition - JMP dialog box shown in Section SG17.3):

1. Drag **Upgraded** to the **Y, Response** box.
2. Drag **Purchases** to the **X, Factor** box.
3. Drag **Extra Cards** to the **X, Factor** box.
4. Click **OK**.

In the next dialog box:

5. Leave **Holdback** as the **Validation Method** and **0.3333** as the **Holdback Proportion**.
6. Enter **2** as the number of **Hidden Nodes**.
7. Click **Go**.

In the DATA - Neural of Upgraded - JMP window:

8. Click the triangle icons to the left of the two **Confusion Matrix** titles and the two **Confusion Rates** titles to display these tables.
9. Click the drop-down button to the left of the title **Model NTanH(2)** and click **Show Estimates**.

Because the initial weights for the model are chosen at random, the actual results will almost certainly differ from the Figure 17.13 results. If the validation data misclassification rate is too high, you may generate additional models. To generate an additional model, click the triangle icon to the left of the title **Model Launch** and then click **Go**. To eliminate a model generated, right-click on the model's **Model NTanH(2)** title and click **Remove Fit** in the shortcut menu.

To perform the Figure 17.14 MLP analysis for predicting the sales of OmniPower bars, open the **OmniPower Excel workbook**. Select **File → Open**. In the Open Data File dialog box, navigate to the location of the file, select the file, and then click **Open**. Because JMP properly identifies Sales, Price, and Promotion as numerical variables, there is no need to change variable types as is done in the preceding example. In the DATA - JMP window, select **Analyze → Modeling → Neural**. In the Neural JMP dialog box:

1. Drag **Sales** to the **Y, Response** box.
2. Drag **Price** to the **X, Factor** box.
3. Drag **Promotion** to the **X, Factor** box.
4. Click **OK**.

In the next dialog box:

5. Leave **Holdback** as the **Validation Method** and **0.3333** as the **Holdback Proportion**.
6. Enter **3** as the number of **Hidden Nodes**.
7. Click **Go**.

In the DATA - Neural of Sales - JMP window:

8. Click the drop-down button to the left of the title **Model NTanH(3)** and click **Show Estimates**.

As noted in the first example, because the initial weights for the model are chosen at random, the actual results will almost certainly differ from the Figure 17.14 results. If necessary, you can generate additional models by clicking **Go** in **Model Launch** as explained in the previous example.

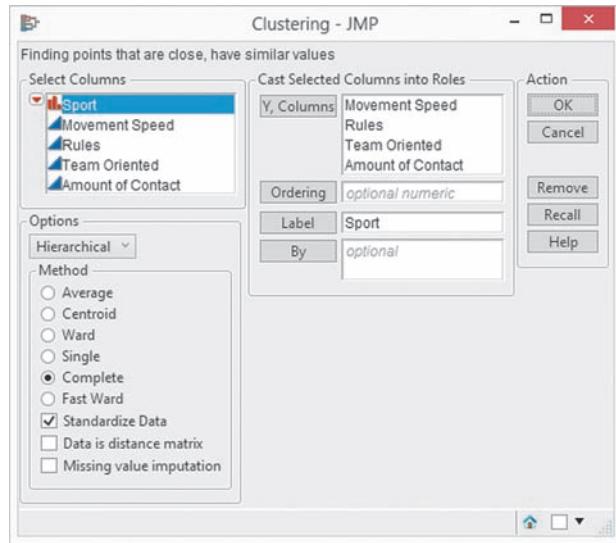
SG17.5 CLUSTER ANALYSIS

JMP 10 Use Cluster.

For example, to perform the Figure 17.15 cluster analysis for the different sports, open the **Sports Excel workbook**. Select **File → Open**. In the Open Data File dialog box, navigate to the location of the file, select the file, and then click **Open**. (JMP properly identifies all five variables that comprise the file.) In the DATA - JMP window, select **Analyze → Multivariate Methods → Cluster**. In the Cluster dialog box (shown in the right column):

1. Drag **Movement Speed** to the **Y, Columns** box.
2. Drag **Rules** to the **Y, Columns** box.
3. Drag **Team Oriented** to the **Y, Columns** box.

4. Drag **Amount of Contact** to the **Y, Columns** box.
5. Drag **Sport** to the **Label** box.
6. In the **Options drop-down list**, click **Hierarchical** and click **Complete** (in the **Method** group).
7. Click **OK**.



In the DATA - Hierarchical Cluster - JMP dialog box, click the triangle icon to the left of the title **Clustering History** to reveal how the clustering was done.

SG17.6 MULTIDIMENSIONAL SCALING

JMP Use the Multidimensional Scaling add-in and the R statistical package

For example, to perform the Figure 17.16 multidimensional scaling analysis based on the mean score of each attribute for each sport, open the **Sports Excel workbook**. Select **File → Open**. In the Open Data File dialog box, navigate to the location of the file, select the file, and then click **Open**. (JMP properly identifies all five variables that comprise the file.) In the DATA - JMP window, click **Add-Ins → Multidimensional Scaling**. In the General tab of the Multidimensional Scaling - JMP dialog box:

1. Click **R project**.
2. Drag **Movement Speed** to the **Y, Columns** box.
3. Drag **Rules** to the **Y, Columns** box.
4. Drag **Team Oriented** to the **Y, Columns** box.
5. Drag **Sport** to the **Label Column** box.
6. Click **Run**.

The add-in displays an overlay plot in a new NMDS Fit Output - JMP window. To view the results for a particular number of dimensions, click the plot point for that number of dimensions. Then click **Display MDS Results for Selected Dimension**. To view the table of stress values, click the drop-down list button to the left of the title **Overlay Plot** and click **Script**, then **Data Table Window**.

The add-in also works with SAS. To use SAS, click **SAS Project** in step 1.

CHAPTER 18

A Roadmap for Analyzing Data

CONTENTS

18.1 Analyzing Numerical Variables

Describing the Characteristics of a Numerical Variable

Reaching Conclusions About the Population Mean and/or Standard Deviation

Determining Whether the Mean and/or Standard Deviation Differs Depending on the Group

Determining Which Factors Affect the Value of a Variable

Predicting the Value of a Variable Based on the Values of Other Variables

Determining Whether the Values of a Variable Are Stable over Time

18.2 Analyzing Categorical Variables

Describing the Proportion of Items of Interest in Each Category

Reaching Conclusions About the Proportion of Items of Interest

Determining Whether the Proportion of Items of Interest Differs Depending on the Group

Predicting the Proportion of Items of Interest Based on the Values of Other Variables

Determining Whether the Proportion of Items of Interest Is Stable over Time

USING STATISTICS: Mounting Future Analyses, Revisited

OBJECTIVES

Identify the questions to ask when choosing which statistical methods to use to conduct data analysis

Generate rules for applying statistics in future studies and analyses

USING STATISTICS

Mounting Future Analyses

Learning business statistics is a lot like climbing a mountain. At first, it may seem intimidating, or even overwhelming, but over time you learn techniques that help make the task much more manageable. In Section GS.1, you learned how the DCOVA framework can make the big task of applying statistics to business problems more manageable. After learning methods in early chapters to Define, Collect, and Organize data, you have spent most of your time studying ways to Visualize and Analyze data.

Determining what methods to use to analyze data may have seemed straightforward when doing homework problems from a particular chapter, but what approach will you take when you find yourself in new situations, needing to analyze

data for another course or to help solve a problem in a real business setting? After all, when you solved a problem from a chapter on multiple regression, you “knew” that multiple regression methods would be part of your analysis. In new situations, you might wonder whether you should use multiple regression—or whether using simple linear regression would be better—or whether *any* type of regression would be appropriate. You also might wonder if you should use a combination of methods from several different chapters to help solve the problems you face.

The question for you becomes: How can you apply the statistical methods you have learned to new situations that require you to analyze data?



Courtesy of Sharyn Rosenberg

Reviewing Table 18.1, which contains a summary of the contents of this book, arranged by data analysis task, would be a good starting point for answering the question posed in the Using Statistics scenario.

TABLE 18.1

Commonly Used
Data Analysis Tasks
Discussed in This Book

DESCRIBING A GROUP OR SEVERAL GROUPS	
For Numerical Variables:	Ordered array, stem-and-leaf display, frequency distribution, relative frequency distribution, percentage distribution, cumulative percentage distribution, histogram, polygon, cumulative percentage polygon (Sections 2.2 and 2.4)
Boxplot (Section 3.3)	Normal probability plot (Section 6.3)
Bullet graph, gauge, treemap (Section 17.1)	Mean, median, mode, quartiles, geometric mean, range, interquartile range, standard deviation, variance, coefficient of variation, skewness, kurtosis (Sections 3.1, 3.2, and 3.3)
Index numbers (online Section 16.8)	
For Categorical Variables:	Summary table, bar chart, pie chart, Pareto chart (Sections 2.1 and 2.3)
	Contingency tables and multidimensional tables (Sections 2.1 and 2.6)
MAKING INFERENCES ABOUT ONE GROUP	
For Numerical Variables:	Confidence interval estimate of the mean (Sections 8.1 and 8.2)
	<i>t</i> test for the mean (Section 9.2)
	Chi-square test for a variance or standard deviation (online Section 12.7)
For Categorical Variables:	Confidence interval estimate of the proportion (Section 8.3)
	Z test for the proportion (Section 9.4)
COMPARING TWO GROUPS	
For Numerical Variables:	Tests for the difference in the means of two independent populations (Section 10.1)
	Wilcoxon rank sum test (Section 12.4)
	Paired <i>t</i> test (Section 10.2)
	<i>F</i> test for the difference between two variances (Section 10.4)
	Wilcoxon signed rank test (online Section 12.8)
For Categorical Variables:	Z test for the difference between two proportions (Section 10.3)
	Chi-square test for the difference between two proportions (Section 12.1)
	McNemar test for two related samples (online Section 12.6)
COMPARING MORE THAN TWO GROUPS	
For Numerical Variables:	One-way analysis of variance (Section 11.1)
	Kruskal-Wallis rank test (Section 12.5)
	Randomized block design (Section 11.2)
	Two-way analysis of variance (Section 11.3)
	Friedman rank test (online Section 12.9)
For Categorical Variables:	Chi-square test for differences among more than two proportions (Section 12.2)
ANALYZING THE RELATIONSHIP BETWEEN TWO VARIABLES	
For Numerical Variables:	Scatter plot, time-series plot (Section 2.5)
	Covariance, coefficient of correlation, <i>t</i> test of correlation (Sections 3.5 and 13.7)
	Simple linear regression (Chapter 13)
	Time-series forecasting (Chapter 16)
For Categorical Variables:	Contingency table, side-by-side bar chart (Sections 2.1 and 2.3)
	Chi-square test of independence (Section 12.3)

ANALYZING THE RELATIONSHIP BETWEEN TWO OR MORE VARIABLES**For Numerical Dependent Variables:**

- Multiple regression (**Chapters 14 and 15**)
- Regression tree (**Section 17.3**)
- Neural network (**Section 17.4**)

For Categorical Dependent Variables:

- Logistic regression (**Section 14.7**)
- Classification tree (**Section 17.3**)
- Neural network (**Section 17.4**)

CLASSIFYING OBJECTS INTO GROUPS**For a Set of Variables:**

- Cluster analysis (**Section 17.5**)
- Multidimensional scaling (**Section 17.6**)

ANALYZING PROCESS DATA**For Numerical Variables:**

- \bar{X} and R control charts (**online Section 19.5**)

For Categorical Variables:

- p chart (**online Section 19.2**)

For Counts of Nonconformities:

- c chart (**online Section 19.4**)

 **Student Tip**

Recall that *numerical variables* have values that represent quantities, while *categorical variables* have values that can only be placed into categories, such as yes and no.

In the DCOVA approach, the first thing you do is to *define* the variables that you want to study in order to solve a business problem or meet a business objective. To do this, you must identify the type of business problem (whether you are describing a group or making inferences about a group, among other choices) and then determine the type of variable—numerical or categorical—you are analyzing.

In Table 18.1, the all-uppercase first-level headings identify types of business problems, and the second-level headings always include the two types of variables. The entries in Table 18.1 identify the specific statistical methods appropriate for a particular type of business problem and type of variable.

Choosing appropriate statistical methods for your data is the single most important task you face and is at the heart of “doing statistics.” But this selection process is also the single most difficult thing you do when applying statistics! How, then, can you ensure that you have made an appropriate choice? By asking a series of questions, you can guide yourself to the appropriate choice of methods.

The rest of this chapter presents questions that will help guide you in making this choice. Two lists of questions, one for numerical variables and the other for categorical variables, are presented in the next two sections. Having two lists makes the decision you face more manageable while also reinforcing the importance of identifying the type of variable that you seek to analyze.

18.1 Analyzing Numerical Variables

Exhibit 18.1 presents the list of questions to ask if you plan to analyze a numerical variable. Each question is independent of the others, and you can ask as many or as few questions as is appropriate for your analysis. How to go about answering these questions follows Exhibit 18.1.

EXHIBIT 18.1 Questions to Ask When Analyzing Numerical Variables

- Do you want to describe the characteristics of the variable (possibly broken down into several groups)?
- Do you want to reach conclusions about the mean and/or standard deviation of the variable in a population?
- Do you want to determine whether the mean and/or standard deviation of the variable differs depending on the group?
- Do you want to determine which factors affect the value of a variable?
- Do you want to predict the value of the variable based on the values of other variables?
- Do you want to determine whether the values of the variable are stable over time?

Describing the Characteristics of a Numerical Variable

You develop tables and charts and compute descriptive statistics to describe characteristics such as central tendency, variation, and shape. Specifically, you can create a stem-and-leaf display, percentage distribution, histogram, polygon, boxplot, normal probability plot, bullet graph, gauge, and treemap (see Sections 2.2, 2.4, 3.3, 6.3, and 17.1), and you can compute statistics such as the mean, median, mode, quartiles, range, interquartile range, standard deviation, variance, coefficient of variation, skewness, and kurtosis (see Sections 3.1, 3.2, and 3.3).

Reaching Conclusions About the Population Mean and/or Standard Deviation

You have several different choices, and you can use any combination of these choices. To estimate the mean value of the variable in a population, you construct a confidence interval estimate of the mean (see Section 8.2). To determine whether the population mean is equal to a specific value, you conduct a t test of hypothesis for the mean (see Section 9.2). To determine whether the population standard deviation or variance is equal to a specific value, you conduct a χ^2 test of hypothesis for the standard deviation or variance (see online Section 12.7).

Determining Whether the Mean and/or Standard Deviation Differs Depending on the Group

When examining differences between groups, you first need to establish which categorical variable to use to divide your data into groups. You then need to know whether this grouping variable divides your data in two groups (such as male and female groups for a gender variable) or whether the variable divides your data into more than two groups (such as the four in-store locations for mobile electronics discussed in Section 11.1). Finally, you must ask whether your data set contains independent groups or whether your data set contains matched or repeated measurements.

If the Grouping Variable Defines Two Independent Groups and You Are Interested in Central Tendency

Which hypothesis tests you use depends on the assumptions you make about your data.

If you assume that your numerical variable is normally distributed and that the variances are equal, you conduct a pooled t test for the difference between the means (see Section 10.1). If you cannot assume that the variances are equal, you conduct a separate-variance t test for the difference between the means (see Section 10.1). To test whether the variances are equal, assuming that the populations are normally distributed, you can conduct an F test for the differences between the variances. In either case, if you believe that your numerical variables are not normally distributed, you can perform a Wilcoxon rank sum test (see Section 12.4) and compare the results of this test to those of the t test.

To evaluate the assumption of normality that the pooled t test and separate-variance t test include, you can construct boxplots and normal probability plots for each group.

If the Grouping Variable Defines Two Groups of Matched Samples or Repeated Measurements and You Are Interested in Central Tendency

If you can assume that the paired differences are normally distributed, you conduct a paired t test (see Section 10.2). If you cannot assume that the paired differences are normally distributed, you conduct a Wilcoxon signed rank test (see online Section 12.8).

If the Grouping Variable Defines Two Independent Groups and You Are Interested in Variability

If you can assume that your numerical variable is normally distributed, you conduct an F test for the difference between two variances (see Section 10.4).

If the Grouping Variable Defines More Than Two Independent Groups and You Are Interested in Central Tendency

If you can assume that the values of the numerical variable are normally distributed, you conduct a one-way analysis of variance (see Section 11.1); otherwise, you conduct a Kruskal-Wallis rank test (see Section 12.5).

If the Grouping Variable Defines More Than Two Groups of Matched Samples or Repeated Measurements and You Are Interested in Central Tendency Suppose that you have a design where the rows represent the blocks and the columns represent the levels of a factor. If you can assume that the values of the numerical variable are normally distributed, you conduct a randomized block design *F* test (see Section 11.2). If you cannot assume that the paired differences are normally distributed, you conduct a Friedman rank test (see online Section 12.9)

Determining Which Factors Affect the Value of a Variable

If there are two factors to be examined to determine their effect on the values of a variable, you develop a two-factor factorial design (see Section 11.3).

Predicting the Value of a Variable Based on the Values of Other Variables

When predicting the values of a numerical dependent variable, you conduct least-squares regression analysis. The least-squares regression model you develop depends on the number of independent variables in your model. If there is only one independent variable being used to predict the numerical dependent variable of interest, you develop a simple linear regression model (see Chapter 13); otherwise, you develop a multiple regression model (see Chapters 14 and 15) or a regression tree (see Section 17.3) or a neural network (see Section 17.4).

If you have values over a period of time and you want to forecast the variable for future time periods, you can use moving averages, exponential smoothing, least-squares forecasting, and autoregressive modeling (see Chapter 16).

Determining Whether the Values of a Variable Are Stable Over Time

If you are studying a process and have collected data on the values of a numerical variable over a time period, you construct *R* and \bar{X} charts (see online Section 19.5). If you have collected data in which the values are counts of the number of nonconformities, you construct a *c* chart (see online Section 19.4).

18.2 Analyzing Categorical Variables

Exhibit 18.2 presents the list of questions to ask if you plan to analyze a categorical variable. Each question is independent of the others, and you can ask as many or as few questions as is appropriate for your analysis. How to go about answering these questions follows Exhibit 18.2.

EXHIBIT 18.2

Questions to Ask When Analyzing Categorical Variables

- Do you want to describe the proportion of items of interest in each category (possibly broken down into several groups)?
- Do you want to reach conclusions about the proportion of items of interest in a population?
- Do you want to determine whether the proportion of items of interest differs depending on the group?
- Do you want to predict the proportion of items of interest based on the values of other variables?
- Do you want to determine whether the proportion of items of interest is stable over time?

Describing the Proportion of Items of Interest in Each Category

You create summary tables and use these charts: bar chart, pie chart, Pareto chart, or side-by-side bar chart (see Sections 2.1 and 2.3).

Reaching Conclusions About the Proportion of Items of Interest

You have two different choices. You can estimate the proportion of items of interest in a population by constructing a confidence interval estimate of the proportion (see Section 8.3). Or, you can determine whether the population proportion is equal to a specific value by conducting a Z test of hypothesis for the proportion (see Section 9.4).

Determining Whether the Proportion of Items of Interest Differs Depending on the Group

When examining this difference, you first need to establish the number of categories associated with your categorical variable and the number of groups in your analysis. If your data contain two groups, you must also ask if your data contain independent groups or if your data contain matched samples or repeated measurements.

For Two Categories and Two Independent Groups You conduct either the Z test for the difference between two proportions (see Section 10.3) or the χ^2 test for the difference between two proportions (see Section 12.1).

For Two Categories and Two Groups of Matched or Repeated Measurements You conduct the McNemar test (see online Section 12.6).

For Two Categories and More Than Two Independent Groups You conduct a χ^2 test for the difference among several proportions (see Section 12.2).

For More Than Two Categories and More Than Two Groups You develop contingency tables and use multidimensional contingency tables to drill down to examine relationships among two or more categorical variables (Sections 2.1, 2.6, and 17.1). When you have two categorical variables, you conduct a χ^2 test of independence (see Section 12.3).

Predicting the Proportion of Items of Interest Based on the Values of Other Variables

You develop a logistic regression model (see Section 14.7) or you use classification trees (see Section 17.3) or neural networks (see Section 17.4).

Determining Whether the Proportion of Items of Interest Is Stable Over Time

If you are studying a process and have collected data over a time period, you can create the appropriate control chart. If you have collected the proportion of items of interest over a time period, you develop a p chart (see online Section 19.2).

USING STATISTICS

Mounting Future Analyses, Revisited

This chapter summarizes all the methods discussed in the first 17 chapters of this book. The data analysis methods discussed in the book are organized in Table 18.1 according to whether each method is used for describing a group or several groups, for making inferences about one group or comparing two or more groups, or for analyzing relationships between two or more variables. Then, sets of questions

are listed in Exhibits 18.1 and 18.2 to assist you in determining what method to use to analyze your data.



Courtesy of Sharyn Rosenberg

Digital Case

Whereas other Digital Cases asked you to apply your knowledge about the proper use of statistics, this case helps you remember how to properly apply that knowledge.

Guadalupe Cooper and Gilbert Chandler had worked very hard all semester long in their business statistics course. They now faced a final project in which they had to establish a plan to analyze a set of data that had been assigned to

them by their instructor. As they looked through the online materials for their statistics textbook, they found **DataAnalysisGuide.pdf** in the Digital Case materials. “Gee, this is like the material in Chapter 18, but in interactive form!” one of them noted. They both then knew what questions they needed to ask in order to get started on their final semester task.

CHAPTER REVIEW PROBLEMS

18.1 In many manufacturing processes, the term *work-in-process* (often abbreviated WIP) is used. At the LSS Publishing book manufacturing plants, WIP represents the time it takes for sheets from a press to be folded, gathered, sewn, tipped on end sheets, and bound together to form a book, and the book placed in a packing carton. The operational definition of the variable of interest, processing time, is the number of days (measured in hundredths) from when the sheets come off the press to when the book is placed in a packing carton. The company has the business objective of determining whether there are differences in the WIP between plants. Data have been collected from samples of 20 books at each of two production plants. The data, stored in **WIP**, are as follows:

Plant A

5.62	5.29	16.25	10.92	11.46	21.62	8.45	8.58	5.41	11.42
11.62	7.29	7.50	7.96	4.42	10.50	7.58	9.29	7.54	8.92

Plant B

9.54	11.46	16.62	12.62	25.75	15.41	14.29	13.13	13.71	10.04
5.75	12.46	9.17	13.21	6.00	2.33	14.25	5.37	6.25	9.71

Completely analyze the data.

18.2 Many factors determine the attendance at Major League Baseball games. These factors can include when the game is played, the weather, the opponent, whether the team is having a good season, and whether a marketing promotion is held. Popular promotions during a recent season included the traditional hat days and poster days and the newer craze, bobble-heads of star players.

(Data extracted from T. C. Boyd and T. C. Krehbiel, “An Analysis of the Effects of Specific Promotion Types on Attendance at Major League Baseball Games,” *Mid-American Journal of Business*, 2006, 21, pp. 21–32.) The file **Baseball** includes the following variables for a recent Major League Baseball season:

TEAM—Kansas City Royals, Philadelphia Phillies,
Chicago Cubs, or Cincinnati Reds

ATTENDANCE—Paid attendance for the game

TEMP—High temperature for the day

WIN%—Team’s winning percentage at the time of the game

OPWIN%—Opponent team’s winning percentage at the time of the game

WEEKEND—1 if game played on Friday, Saturday, or Sunday; 0 otherwise

PROMOTION—1 if a promotion was held; 0 if no promotion was held

You want to predict attendance and determine the factors that influence attendance. Completely analyze the data for the Kansas City Royals.

18.3 Repeat Problem 18.2 for the Philadelphia Phillies.

18.4 Repeat Problem 18.2 for the Chicago Cubs.

18.5 Repeat Problem 18.2 for the Cincinnati Reds.

18.6 The file **EuroTourism2** contains a sample of 27 European countries. Variables included are the number of jobs generated in the travel and tourism industry in 2012, the spending on business travel

within the country by residents and international visitors in 2012, the total number of international visitors who visited the country in 2012, and the number of establishments that provide overnight accommodation for tourists. (Data extracted from www.marketline.com.) You want to be able to predict the number of jobs generated in the travel and tourism industry. Completely analyze the data.

18.7 The file **Philly** contains a sample of 25 neighborhoods in Philadelphia. Variables included are neighborhood population, median sales price of homes in 2012, mean number of days homes were on the market in 2012, number of homes sold in 2012, median neighborhood household income, percentage of residents in the neighborhood with a bachelor's degree or higher, and whether the neighborhood is considered "hot" (coded as 1 = yes, 0 = no). (Data extracted from bit.ly/13M7KuP.) You want to be able to predict median sales price of homes. Completely analyze the data.

18.8 Professional basketball has truly become a sport that generates interest among fans around the world. More and more players come from outside the United States to play in the National Basketball Association (NBA). Many factors could impact the number of wins achieved by each NBA team. In addition to the number of wins, the file **NBA2012** contains team statistics for points per game (for team, opponent, and the difference between team and opponent), field goal (shots made) percentage (for team, opponent, and the difference between team and opponent), turnovers (losing the ball before a shot is taken) per game (for team, opponent, and the difference between team and opponent), and the rebound percentage. You want to be able to predict the number of wins. Completely analyze the data.

18.9 The data in **UsedCars** represent characteristics of cars that are currently part of an inventory of a used car dealership. The variables included are car, year, age, price (\$), mileage, power (hp), and fuel (mpg).

You want to describe each of these variables, and you would like to predict the price of the used cars. Analyze the data.

18.10 A study was conducted to determine whether any gender bias existed in an academic science environment. Faculty from several universities were asked to rate candidates for the position of undergraduate laboratory manager based on their application. The gender of the applicant was given in the applicant's materials. The raters were from either biology, chemistry, or physics departments. Each rater was to give a competence rating to the applicant's materials on a seven point scale with 1 being the lowest and 7 being the highest. In addition, the rater supplied a starting salary that should be offered to the applicant. These data (which have been altered from an actual study to preserve the anonymity of the respondents) are stored in **Candidate Assessment**.

Analyze the data. Do you think that there is any gender bias in the evaluations? Support your point of view with specific references to your data analysis.

18.11 Zagat's publishes restaurant ratings for various locations in the United States. The file **Restaurants2** contains the Zagat rating for food, décor, service, cost per person, and popularity index (popularity points the restaurant received divided by the number of people who voted for that restaurant) for various types of restaurants in New York City.

You want to study differences in the cost of a meal for the different types of cuisines and also want to be able to predict the cost of a meal. Completely analyze the data. (Data extracted from *Zagat Survey 2012 New York City Restaurants*).

18.12 The data in the file **BankMarketing** are from a direct marketing campaign conducted by a Portuguese banking institution (Data extracted from S. Moro, R. Laureano and P. Cortez, "Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology," in P. Novais et al. (Eds.), *Proceedings of the European Simulation and Modeling Conference—ESM'2011*, pp. 117–121.) The variables included were age, type of job, marital status, education, whether credit is in default, average yearly balance in account in Euros, whether there is a housing loan, whether there is a personal loan, last contact duration in seconds, number of contacts performed during this campaign, and has the client purchased a term deposit.

Analyze the data and assess the likelihood that the client will purchase a term deposit.

18.13 A mining company operates a large heap-leach gold mine in the western United States. The gold mined at this location consists of ore that is very low grade, having about 0.0032 ounce of gold in 1 ton of ore. The process of heap-leaching involves the mining, crushing, stacking, and leaching millions of tons of gold ore per year. In the process, ore is placed in a large heap on an impermeable pad. A weak chemical solution is sprinkled over the heap and is collected at the bottom after percolating through the ore. As the solution percolates through the ore, the gold is dissolved and is later recovered from the solution. This technology, which has been used for more than 30 years, has made the operation profitable. Due to the large amount of ore that is handled, the company is continually exploring ways to improve the process. As part of an expansion several years ago, the stacking process was automated with the construction of a computer controlled stacker. This stacker was designed to load 35,000 tons of ore per day at a cost that was less than the previous process that used manually operated trucks and bulldozers. However, since its installation, the stacker has not been able to achieve these results consistently. Data for a recent 35-day period that indicate the amount stacked (tons) and the downtime (minutes) are stored in the file **Mining**. Other data that indicate the causes for the downtime are stored in **Mining2**.

Analyze the data, making sure to present conclusions about the daily amount stacked and the causes of the downtime. In addition, be sure to develop a model to predict the amount stacked based on downtime.

18.14 A survey was conducted on the characteristics of households in the United States. The data (which have been altered from an actual study to preserve the anonymity of the respondents) are stored in **Households**. The variables are gender, age, Hispanic origin, type of dwelling, age of dwelling in years, years living at dwelling, number of bedrooms, number of vehicles kept at dwelling, fuel type at dwelling, monthly cost of fuel at dwelling (\$), U.S. citizenship, college degree, marital status, work for pay in previous week, mode of transportation to work, commuting time in minutes, hours worked per week, type of organization, annual earned income (\$), and total annual income (\$).

Analyze these data and prepare a report describing your conclusions.

18.15 The file **HybridSales** contains the number of domestic and imported hybrid vehicles sold in the United States from 1999 to 2012. (Data extracted from bit.ly/17hJk9H.) You want to be able to predict the number of domestic and imported hybrid vehicles sold in the United States in 2013 and 2014. Completely analyze the data.

Appendices

A. BASIC MATH CONCEPTS AND SYMBOLS

- A.1** Rules for Arithmetic Operations
- A.2** Rules for Algebra: Exponents and Square Roots
- A.3** Rules for Logarithms
- A.4** Summation Notation
- A.5** Statistical Symbols
- A.6** Greek Alphabet

B. REQUIRED EXCEL SKILLS

- B.1** Worksheet Entries and References
- B.2** Absolute and Relative Cell References
- B.3** Entering Formulas into Worksheets
- B.4** Pasting with Paste Special
- B.5** Basic Worksheet Formatting
- B.6** Chart Formatting
- B.7** Selecting Cell Ranges for Charts
- B.8** Deleting the "Extra" Bar from a Histogram
- B.9** Creating Histograms for Discrete Probability Distributions

C. ONLINE RESOURCES

- C.1** About the Online Resources for This Book
- C.2** Accessing the MyStatLab Course Online
- C.3** Details of Downloadable Files
- C.4** PHStat

D. CONFIGURING MICROSOFT EXCEL

- D.1** Getting Microsoft Excel Ready for Use (ALL)
- D.2** Getting PHStat Ready for Use (ALL)
- D.3** Configuring Excel Security for Add-In Usage (WIN)
- D.4** Opening PHStat (ALL)

- D.5** Using a Visual Explorations Add-in Workbook (ALL)

- D.6** Checking for the Presence of the Analysis ToolPak or Solver Add-Ins (ALL)

E. TABLES

- E.1** Table of Random Numbers
- E.2** The Cumulative Standardized Normal Distribution
- E.3** Critical Values of t
- E.4** Critical Values of χ^2
- E.5** Critical Values of F
- E.6** Lower and Upper Critical Values, T_1 , of the Wilcoxon Rank Sum Test
- E.7** Critical Values of the Studentized Range, Q
- E.8** Critical Values, d_L and d_U , of the Durbin-Watson Statistic, D
- E.9** Control Chart Factors
- E.10** The Standardized Normal Distribution

F. USEFUL EXCEL KNOWLEDGE

- F.1** Useful Keyboard Shortcuts
- F.2** Verifying Formulas and Worksheets
- F.3** New Function Names
- F.4** Understanding the Nonstatistical Functions

G. SOFTWARE FAQS

- G.1** PHStat FAQs
- G.2** Microsoft Excel FAQs
- G.3** FAQs for New Microsoft Excel 2013 Users
- G.4** Minitab FAQs

SELF-TEST SOLUTIONS AND ANSWERS TO SELECTED EVEN-NUMBERED PROBLEMS

A.1 Rules for Arithmetic Operations

RULE	EXAMPLE
1. $a + b = c$ and $b + a = c$	$2 + 1 = 3$ and $1 + 2 = 3$
2. $a + (b + c) = (a + b) + c$	$5 + (7 + 4) = (5 + 7) + 4 = 16$
3. $a - b = c$ but $b - a \neq c$	$9 - 7 = 2$ but $7 - 9 \neq 2$
4. $(a)(b) = (b)(a)$	$(7)(6) = (6)(7) = 42$
5. $(a)(b + c) = ab + ac$	$(2)(3 + 5) = (2)(3) + (2)(5) = 16$
6. $a \div b \neq b \div a$	$12 \div 3 \neq 3 \div 12$
7. $\frac{a + b}{c} = \frac{a}{c} + \frac{b}{c}$	$\frac{7 + 3}{2} = \frac{7}{2} + \frac{3}{2} = 5$
8. $\frac{a}{b + c} \neq \frac{a}{b} + \frac{a}{c}$	$\frac{3}{4 + 5} \neq \frac{3}{4} + \frac{3}{5}$
9. $\frac{1}{a} + \frac{1}{b} = \frac{b + a}{ab}$	$\frac{1}{3} + \frac{1}{5} = \frac{5 + 3}{(3)(5)} = \frac{8}{15}$
10. $\left(\frac{a}{b}\right)\left(\frac{c}{d}\right) = \left(\frac{ac}{bd}\right)$	$\left(\frac{2}{3}\right)\left(\frac{6}{7}\right) = \left(\frac{(2)(6)}{(3)(7)}\right) = \frac{12}{21}$
11. $\frac{a}{b} \div \frac{c}{d} = \frac{ad}{bc}$	$\frac{5}{8} \div \frac{3}{7} = \left(\frac{(5)(7)}{(8)(3)}\right) = \frac{35}{24}$

A.2 Rules for Algebra: Exponents and Square Roots

RULE	EXAMPLE
1. $(X^a)(X^b) = X^{a+b}$	$(4^2)(4^3) = 4^5$
2. $(X^a)^b = X^{ab}$	$(2^2)^3 = 2^6$
3. $(X^a/X^b) = X^{a-b}$	$\frac{3^5}{3^3} = 3^2$
4. $\frac{X^a}{X^a} = X^0 = 1$	$\frac{3^4}{3^4} = 3^0 = 1$
5. $\sqrt{XY} = \sqrt{X}\sqrt{Y}$	$\sqrt{(25)(4)} = \sqrt{25}\sqrt{4} = 10$
6. $\sqrt{\frac{X}{Y}} = \frac{\sqrt{X}}{\sqrt{Y}}$	$\sqrt{\frac{16}{100}} = \frac{\sqrt{16}}{\sqrt{100}} = 0.40$

A.3 Rules for Logarithms

Base 10

\log is the symbol used for base-10 logarithms:

RULE	EXAMPLE
1. $\log(10^a) = a$	$\log(100) = \log(10^2) = 2$
2. If $\log(a) = b$, then $a = 10^b$	If $\log(a) = 2$, then $a = 10^2 = 100$
3. $\log(ab) = \log(a) + \log(b)$	$\log(100) = \log[(10)(10)] = \log(10) + \log(10)$ $= 1 + 1 = 2$
4. $\log(a^b) = (b) \log(a)$	$\log(1,000) = \log(10^3) = (3) \log(10) = (3)(1) = 3$
5. $\log(a/b) = \log(a) - \log(b)$	$\log(100) = \log(1,000/10) = \log(1,000) - \log(10)$ $= 3 - 1 = 2$

EXAMPLE

Take the base-10 logarithm of each side of the following equation:

$$Y = \beta_0 \beta_1^X \varepsilon$$

SOLUTION: Apply rules 3 and 4:

$$\begin{aligned} \log(Y) &= \log(\beta_0 \beta_1^X \varepsilon) \\ &= \log(\beta_0) + \log(\beta_1^X) + \log(\varepsilon) \\ &= \log(\beta_0) + X \log(\beta_1) + \log(\varepsilon) \end{aligned}$$

Base e

\ln is the symbol used for base e logarithms, commonly referred to as natural logarithms. e is Euler's number, and $e \approx 2.718282$:

RULE	EXAMPLE
1. $\ln(e^a) = a$	$\ln(7.389056) = \ln(e^2) = 2$
2. If $\ln(a) = b$, then $a = e^b$	If $\ln(a) = 2$, then $a = e^2 = 7.389056$
3. $\ln(ab) = \ln(a) + \ln(b)$	$\ln(100) = \ln[(10)(10)]$ $= \ln(10) + \ln(10) = 2.302585 + 2.302585 = 4.605170$
4. $\ln(a^b) = (b) \ln(a)$	$\ln(1,000) = \ln(10^3) = 3 \ln(10) = 3(2.302585) = 6.907755$
5. $\ln(a/b) = \ln(a) - \ln(b)$	$\ln(100) = \ln(1,000/10) = \ln(1,000) - \ln(10)$ $= 6.907755 - 2.302585 = 4.605170$

EXAMPLE

Take the base e logarithm of each side of the following equation:

$$Y = \beta_0 \beta_1^X \varepsilon$$

SOLUTION: Apply rules 3 and 4:

$$\begin{aligned}\ln(Y) &= \ln(\beta_0 \beta_1^X \varepsilon) \\ &= \ln(\beta_0) + \ln(\beta_1^X) + \ln(\varepsilon) \\ &= \ln(\beta_0) + X \ln(\beta_1) + \ln(\varepsilon)\end{aligned}$$

A.4 Summation Notation

The symbol Σ , the Greek capital letter sigma, represents “taking the sum of.” Consider a set of n values for variable X . The expression $\sum_{i=1}^n X_i$ means to take the sum of the n values for variable X . Thus:

$$\sum_{i=1}^n X_i = X_1 + X_2 + X_3 + \cdots + X_n$$

The following problem illustrates the use of the symbol Σ . Consider five values of a variable X : $X_1 = 2$, $X_2 = 0$, $X_3 = -1$, $X_4 = 5$, and $X_5 = 7$. Thus:

$$\sum_{i=1}^5 X_i = X_1 + X_2 + X_3 + X_4 + X_5 = 2 + 0 + (-1) + 5 + 7 = 13$$

In statistics, the squared values of a variable are often summed. Thus:

$$\sum_{i=1}^n X_i^2 = X_1^2 + X_2^2 + X_3^2 + \cdots + X_n^2$$

and, in the example above:

$$\begin{aligned}\sum_{i=1}^5 X_i^2 &= X_1^2 + X_2^2 + X_3^2 + X_4^2 + X_5^2 \\ &= 2^2 + 0^2 + (-1)^2 + 5^2 + 7^2 \\ &= 4 + 0 + 1 + 25 + 49 \\ &= 79\end{aligned}$$

$\sum_{i=1}^n X_i^2$, the summation of the squares, is *not* the same as $\left(\sum_{i=1}^n X_i\right)^2$, the square of the sum:

$$\sum_{i=1}^n X_i^2 \neq \left(\sum_{i=1}^n X_i\right)^2$$

In the example given above, the summation of squares is equal to 79. This is not equal to the square of the sum, which is $13^2 = 169$.

Another frequently used operation involves the summation of the product. Consider two variables, X and Y , each having n values. Then:

$$\sum_{i=1}^n X_i Y_i = X_1 Y_1 + X_2 Y_2 + X_3 Y_3 + \cdots + X_n Y_n$$

Continuing with the previous example, suppose there is a second variable, Y , whose five values are $Y_1 = 1$, $Y_2 = 3$, $Y_3 = -2$, $Y_4 = 4$, and $Y_5 = 3$. Then,

$$\begin{aligned}\sum_{i=1}^n X_i Y_i &= X_1 Y_1 + X_2 Y_2 + X_3 Y_3 + X_4 Y_4 + X_5 Y_5 \\ &= (2)(1) + (0)(3) + (-1)(-2) + (5)(4) + (7)(3) \\ &= 2 + 0 + 2 + 20 + 21 \\ &= 45\end{aligned}$$

In computing $\sum_{i=1}^n X_i Y_i$, you need to realize that the first value of X is multiplied by the first value of Y , the second value of X is multiplied by the second value of Y , and so on. These products are then summed in order to compute the desired result. However, the summation of products is *not* equal to the product of the individual sums:

$$\sum_{i=1}^n X_i Y_i \neq \left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right)$$

In this example,

$$\sum_{i=1}^5 X_i = 13$$

and

$$\sum_{i=1}^5 Y_i = 1 + 3 + (-2) + 4 + 3 = 9$$

so that

$$\left(\sum_{i=1}^5 X_i \right) \left(\sum_{i=1}^5 Y_i \right) = (13)(9) = 117$$

However,

$$\sum_{i=1}^5 X_i Y_i = 45$$

The following table summarizes these results:

VALUE	X_i	Y_i	$X_i Y_i$
1	2	1	2
2	0	3	0
3	-1	-2	2
4	5	4	20
5	7	3	21
	$\sum_{i=1}^5 X_i = 13$	$\sum_{i=1}^5 Y_i = 9$	$\sum_{i=1}^5 X_i Y_i = 45$

Rule 1 The summation of the values of two variables is equal to the sum of the values of each summed variable:

$$\sum_{i=1}^n (X_i + Y_i) = \sum_{i=1}^n X_i + \sum_{i=1}^n Y_i$$

Thus,

$$\begin{aligned}\sum_{i=1}^5(X_i + Y_i) &= (2 + 1) + (0 + 3) + (-1 + (-2)) + (5 + 4) + (7 + 3) \\ &= 3 + 3 + (-3) + 9 + 10 \\ &= 22 \\ \sum_{i=1}^5 X_i + \sum_{i=1}^5 Y_i &= 13 + 9 = 22\end{aligned}$$

Rule 2 The summation of a difference between the values of two variables is equal to the difference between the summed values of the variables:

$$\sum_{i=1}^n(X_i - Y_i) = \sum_{i=1}^n X_i - \sum_{i=1}^n Y_i$$

Thus,

$$\begin{aligned}\sum_{i=1}^5(X_i - Y_i) &= (2 - 1) + (0 - 3) + (-1 - (-2)) + (5 - 4) + (7 - 3) \\ &= 1 + (-3) + 1 + 1 + 4 \\ &= 4 \\ \sum_{i=1}^5 X_i - \sum_{i=1}^5 Y_i &= 13 - 9 = 4\end{aligned}$$

Rule 3 The sum of a constant times a variable is equal to that constant times the sum of the values of the variable:

$$\sum_{i=1}^n cX_i = c \sum_{i=1}^n X_i$$

where c is a constant. Thus, if $c = 2$,

$$\begin{aligned}\sum_{i=1}^5 cX_i &= \sum_{i=1}^5 2X_i = (2)(2) + (2)(0) + (2)(-1) + (2)(5) + (2)(7) \\ &= 4 + 0 + (-2) + 10 + 14 \\ &= 26 \\ c \sum_{i=1}^5 X_i &= 2 \sum_{i=1}^5 X_i = (2)(13) = 26\end{aligned}$$

Rule 4 A constant summed n times will be equal to n times the value of the constant.

$$\sum_{i=1}^n c = nc$$

where c is a constant. Thus, if the constant $c = 2$ is summed 5 times,

$$\begin{aligned}\sum_{i=1}^5 c &= 2 + 2 + 2 + 2 + 2 = 10 \\ nc &= (5)(2) = 10\end{aligned}$$

EXAMPLE

Suppose there are six values for the variables X and Y , such that $X_1 = 2, X_2 = 1, X_3 = 5, X_4 = -3, X_5 = 1, X_6 = -2$ and $Y_1 = 4, Y_2 = 0, Y_3 = -1, Y_4 = 2, Y_5 = 7$, and $Y_6 = -3$. Compute each of the following:

(a) $\sum_{i=1}^6 X_i$

(b) $\sum_{i=1}^6 Y_i$

- (c) $\sum_{i=1}^6 X_i^2$ (g) $\sum_{i=1}^6 (X_i - Y_i)$
 (d) $\sum_{i=1}^6 Y_i^2$ (h) $\sum_{i=1}^6 (X_i - 3Y_i + 2X_i^2)$
 (e) $\sum_{i=1}^6 X_i Y_i$ (i) $\sum_{i=1}^6 (cX_i)$, where $c = -1$
 (f) $\sum_{i=1}^6 (X_i + Y_i)$ (j) $\sum_{i=1}^6 (X_i - 3Y_i + c)$, where $c = +3$

Answers

- (a) 4 (b) 9 (c) 44 (d) 79 (e) 10 (f) (13) (g) -5 (h) 65 (i) -4 (j) -5

REFERENCES

1. Bashaw, W. L., *Mathematics for Statistics* (New York: Wiley, 1969).
2. Lanzer, P., *Basic Math: Fractions, Decimals, Percents* (Hicksville, NY: Video Aided Instruction, 2006).
3. Levine, D. and A. Brandwein, *The MBA Primer: Business Statistics*, 3rd ed. (Cincinnati, OH: Cengage Publishing, 2011).
4. Levine, D., *Statistics* (Hicksville, NY: Video Aided Instruction, 2006).
5. Shane, H., *Algebra I* (Hicksville, NY: Video Aided Instruction, 2006).

A.5 Statistical Symbols

+ add	× multiply
- subtract	÷ divide
= equal to	≠ not equal to
≈ approximately equal to	< less than
> greater than	≤ less than or equal to
≥ greater than or equal to	

A.6 Greek Alphabet

GREEK LETTER	LETTER NAME	ENGLISH EQUIVALENT	GREEK LETTER	LETTER NAME	ENGLISH EQUIVALENT		
A	α	Alpha	a	N	ν	Nu	n
B	β	Beta	b	Ξ	ξ	Xi	x
Γ	γ	Gamma	g	O	ο	Omicron	o
Δ	δ	Delta	d	Π	π	Pi	p
E	ε	Epsilon	ĕ	P	ρ	Rho	r
Z	ζ	Zeta	z	Σ	σ	Sigma	s
H	η	Eta	ĕ	T	τ	Tau	t
Θ	θ	Theta	th	Y	υ	Upsilon	u
I	ι	Iota	i	Φ	ϕ	Phi	ph
K	κ	Kappa	k	X	χ	Chi	ch
Λ	λ	Lambda	l	Ψ	ψ	Psi	ps
M	μ	Mu	m	Ω	ω	Omega	ō

This appendix reviews the Excel skills and operations you need to know in order to make effective use of Microsoft Excel. As stated in Section EG.1 on page 8, if you plan to use the *In-Depth Excel* instructions, you will need to be familiar with the contents of Sections B.1 through B.4 as a minimum. Mastery of the skills and operations in this appendix is less necessary if you plan to use PHStat (or the Analysis ToolPak), but knowing them will prove useful if you need to customize the worksheets that PHStat creates or plan to create your summary presentations from those results.

If you find the level of this appendix too challenging or are unfamiliar with the skills listed in Table EG.A on page 8, you should first download and review “Basic Computing Skills” that is available online (see Appendix C for details).

B.1 Worksheet Entries and References

As discussed in Section GS.4, Microsoft Excel uses worksheets to both store data and present the results of analyses. In worksheet cells, you enter the data for variables, text that serves to label data or title a worksheet, or *formulas*. **Formulas** are instructions that perform a calculation or some other computing task such as logical decision making. Formulas are typically found in worksheets that you use to present intermediate calculations or the results of an analysis. In some cases, formulas create or prepare new data to be analyzed.

Formulas typically use values found in other cells to compute a result that is displayed in the cell that stores the formula. This means that when you see that a particular worksheet cell is displaying the value, say, 5, you cannot determine from casual inspection if the worksheet creator typed the number 5 into the cell or if the creator typed a formula that results in the display of the value 5. This trait of worksheets means you should always carefully review the contents of each worksheet you use. In this book, each worksheet with formulas that you might use is accompanied by a “formulas” worksheet that presents the worksheet in a mode that allows you to see all the formulas that have been entered in the worksheet.

Cell References

Most formulas use values that have been entered into other cells. To refer to those cells, Excel uses an addressing, or *referencing*, system that is based on the tabular nature of a

worksheet. Columns are designated with letters and rows are designated with numbers such that the cell in the first row and first column is called A1, the cell in the third row and first column is called A3, and the cell in the third column and first row is C1. To refer to a cell in a formula, you use a cell reference in the form *WorksheetName!ColumnRow*. For example, Data!A2 refers to the cell in the Data worksheet that is in column A and row 2.

You can also use only the *ColumnRow* portion of a full address—for example, A2—as a shorthand way of referring to a cell that is on the same worksheet as the one into which you are entering a formula. (Excel calls the worksheet into which you are making entries the **current worksheet**.) If the worksheet name contains spaces or special characters, such as **CITY DATA** or **Figure_1.2**, you must enclose the sheet name in a pair of single quotes, as in 'CITY DATA'!A2 or 'Figure_1.2'!A2.

To refer to a group of cells, such as the cells of a column that store the data for a particular variable, you use a cell range. A cell range names the upper-left cell and the lower-right cell of the group, using the form *Worksheet Name!UpperLeftCell:LowerRightCell*. For example, the cell range DATA!A1:A11 identifies the first 11 cells in the first column of the DATA worksheet. Cell ranges can extend over multiple columns; the cell range DATA!A1:D11 would refer to the first 11 cells in the first 4 columns of the worksheet. Cell ranges in the form *Column:Column* (or *Row:Row*) that refer to all cells in a column (or row) are also allowed. In this book, you will occasionally see cell ranges such as B:B that refer to all the cells in a column B for situations in which the number of cell entries in column B would be unknown to the worksheet creator.

As with single cell references, you can skip the *WorksheetName!* part of the reference if you are entering a cell range on the current worksheet. And if a worksheet name contains spaces or special characters, the worksheet name must be enclosed in a pair of single quotes. Note, that in some Excel dialog boxes, you *must* include the worksheet name as part of the cell reference in order to get the proper results. (Such cases are noted in the instructions in this book when they arise.)

Although not used in this book, cell references can include a workbook name in the form '[*WorkbookName*] *WorksheetName!*ColumnRow' or '[*WorkbookName*] *WorksheetName!* UpperLeftCell:LowerRightCell'. You might discover such references if you inadvertently copy certain types of worksheets or chart sheets from one workbook to another.

Recalculation

When you use formulas that refer to other cells, the result displayed by the formulas automatically changes as the values in the cells to which the formula refers change. This process, called **recalculation**, was the original novel feature of worksheet programs and first led to these programs being widely used in accounting.

Recalculation forms the basis for constructing worksheet *templates* and *models*. **Templates** are worksheets in which you only need to enter values to get results. Templates can be reused over and over again, by entering different sets of values. Many of the worksheets illustrated in this book are templates. For those worksheets, you need only to enter new values, typically into cells that are tinted a light turquoise color, to get the results you need. Other worksheets illustrated are **models**, which are similar to templates but require the editing of certain formulas as new values are entered into a worksheet. In this book, worksheet models have been designed to simplify such editing tasks and to provide the most generalized solution.

Worksheets that use formulas capable of recalculation are sometimes called “live” worksheets to distinguish them from worksheets that contain only text and numeric entries (“dead” worksheets). A novel feature of the PHStat add-in that you can use with this book is that just about every worksheet the add-in constructs for you is a “live” worksheet. This means that, as first noted in Section EG.1 on page 8, you get the same results, the same worksheets, whether you use *PHStat* or the *In-Depth Excel* instructions in the Excel Guides. This is dissimilar to many other add-ins that produce results in the form of dead worksheets that cannot be reused in any way.

B.2 Absolute and Relative Cell References

Many worksheets contain columns (or rows) of similar-looking formulas. For example, column C in a worksheet might contain formulas that sum the contents of the column A and column B rows. The formula for cell C2 would be $=A2 + B2$, the formula for cell C3 would be $=A3 + B3$, for cell C4, $=A4 + B4$, and so on, down column C. To avoid the drudgery of typing many similar formulas, you can copy a formula and paste it into all the cells in a selected cell range. For example, to copy a formula that has been entered in cell C2 down the column through row 12:

1. Right-click cell C2 and press **Ctrl+C** to copy the formula. A movie marquee-like highlight appears around cell C2.
2. Select the cell range **C3:C12**.
3. With the cell range highlighted, press **Ctrl+V** to paste the formula into the cells of the cell range.

When you perform this copy-and-paste operation, Excel adjusts these **relative cell references** in formulas so that copying the formula $=A2 + B2$ from cell C2 to cell C3 results in the formula $=A3 + B3$ being pasted into cell C3, the formula $=A4 + B4$ being pasted into cell C4, and so on.

There are circumstances in which you do not want Excel to adjust all or part of a formula. For example, if you were copying the cell C2 formula $=(A2 + B2)/B15$, and cell B15 contained the divisor to be used in all formulas, you would not want to see pasted into cell C3 the formula $=(A3 + B3)/B16$. To prevent Excel from adjusting a cell reference, you use **absolute cell references** by inserting dollar signs (\$) before the column and row references of a relative cell reference. For example, the absolute cell reference $\$B\15 in the copied cell C2 formula $=(A2 + B2)/\$B\15 will cause Excel to paste the formula $=(A2 + B2)/\$B\15 into cell C3.

For ease of reading, formulas shown in the worksheet illustrations in this book show relative cell references, even in cases where using absolute cell references would assist in the physical entry of the formulas. Do not confuse the use of the dollar sign symbol with the worksheet formatting operation that displays numbers as dollar currency amounts. (See Section B.5 to learn how to format cells to display numeric values as dollar currency amounts.)

B.3 Entering Formulas into Worksheets

To enter a formula into a cell, first select the cell and then begin the entry by typing the equal sign (=). What follows the equal sign can be a combination of mathematical and data-processing operations and cell references that is terminated by pressing **Enter**. For simple formulas, you use the symbols +, -, *, /, and ^ for the operations addition, subtraction, multiplication, division, and exponentiation (a number raised to a power), respectively. For example, the formula $=A2 + B2$ adds the contents of cells A2 and B2 displays the sum as the value in the cell containing the formula. To revise a formula, either retype the formula or edit it in the formula bar.

Because formulas display their results and not themselves when entered in a cell, you should always review and verify any formula you enter before you use its worksheet to get results. One way to view all the formulas in a worksheet is to press **Ctrl+`** (grave accent). After your formula review, you can press **Ctrl+`** a second time to restore the normal display of values.

Functions

You can use worksheet functions in formulas to simplify certain arithmetic formulas or to gain access to advanced processing or statistical functions. For example, instead

of typing $=A2 + A3 + A4 + A5 + A6$, you could use the **SUM** function to enter the equivalent, and shorter, formula $=SUM(A2:A6)$. Functions are entered by typing their names followed by a pair of parentheses. For almost all functions, you need to make at least one entry inside the pair of parentheses. For functions that require two or more entries, you separate entries with commas, as in the function **QUARTILE** (*variable cell range, quartile number*) function that is discussed in Section EG3.3.

To use a worksheet function in a formula, either type the function as shown in the instructions in this book or select a function from one of the galleries in the Function Library group of the Formulas tab. For example, to enter the formula $=QUARTILE(A2:A20, 2)$ in cell C2, you could either type these 20 characters directly into the cell or select cell C2 and then select **Formulas** → **More Functions** → **Statistical** and click **QUARTILE** from the drop-down list and then enter **A2:A20** and **2** in the Function Arguments dialog box and click **OK**. (For some functions, the selection process is much shorter and, in Excel versions older than Excel 2007, you select **Formulas** → **Insert Function** and then make the necessary entries and selections in one or more dialog boxes that follow.)

Entering Array Formulas

An **array formula** is a formula that you enter just once but that applies to all of the cells in a selected cell range (the “array”). To enter an array formula, first select the cell range and then type the formula, and then, while holding down the **Ctrl** and **Shift** keys, press **Enter** to enter the array formula into all of the cells of the cell range. (In OS X Excel, you can also press **Command+Enter** to enter an array formula.)

To edit an array formula, you must first select the entire cell range that contains the array formula, then edit the formula and then press **Enter** while holding down **Ctrl+Shift** (or press **Command+Enter**). When you select a cell that contains an array formula, Excel adds a pair of curly braces {} to the display of the formula in the formula bar. These curly braces disappear when you start to edit the formula. Including a pair of curly braces around a formula when documenting a worksheet is a convention to indicate that a particular formula is an array formula, but at no time will you ever type the curly braces when you enter an array formula.

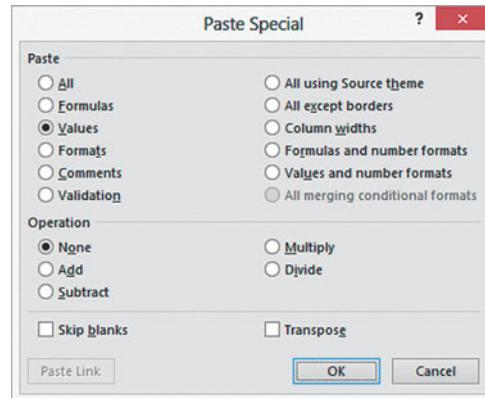
B.4 Pasting with Paste Special

While the keyboard shortcuts **Ctrl+C** and **Ctrl+V** to copy and paste cell contents will often suffice, pasting data from one worksheet to another can sometimes cause unexpected side effects. When the two worksheets are in different workbooks, a simple paste creates an external link to the original workbook. This can lead to errors later if the first workbook is unavailable when the second one is being used. Even pasting

between worksheets in the same workbook can lead to problems if what is being pasted is a cell range of formulas.

To avoid such side effects, use **Paste Special** in such situations. To use this operation, copy the source cell range using **Ctrl+C** and then right-click the cell (or cell range) that is the target of the paste and click **Paste Special** from the shortcut menu.

In the Paste Special dialog box (shown below), click **Values** and then click **OK**. For the first case, Paste Special Values pastes the current values of the cells in the first workbook and not formulas that use cell references to the first workbook.



Paste Special can paste other types of information, including cell formatting information. In some copying contexts, placing the mouse pointer over Paste Special in the shortcut menu will reveal a gallery of shortcuts to the choices presented in the Paste Special dialog box. For a full discussion of these additional features of Paste Special, see the Microsoft Excel help system.

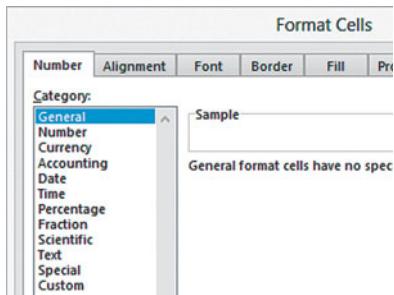
If you use PHStat and have data for a procedure in the form of formulas, copy your data and then use Paste Special to paste columns of equivalent *values*. (Click **Values** in the Paste Special dialog box to create the values.) Then use the columns of values as the cell range of the data for the procedure. PHStat will not work properly if the data for a procedure are in the form of formulas.

B.5 Basic Worksheet Cell Formatting

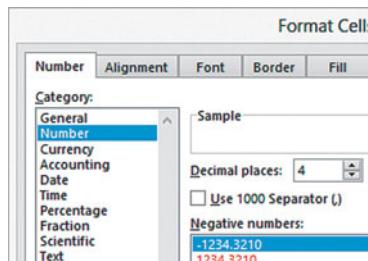
You can change many aspects of how Excel displays the contents of worksheet cells through formatting operations. You format cells either by making entries in the Format Cells dialog box or by clicking shortcut buttons in the Home tab at the top of the Excel window. If you are new to Excel, you may find that using the Format Cells dialog box method to be easier, at least initially. Then, over time, you may want to switch to the Home tab shortcuts discussed later in this section.

To use the Format Cells dialog box, right-click a cell (or cell range) and click **Format Cells** in the shortcut menu.

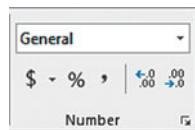
Excel displays the Number tab of the dialog box (partially shown below).



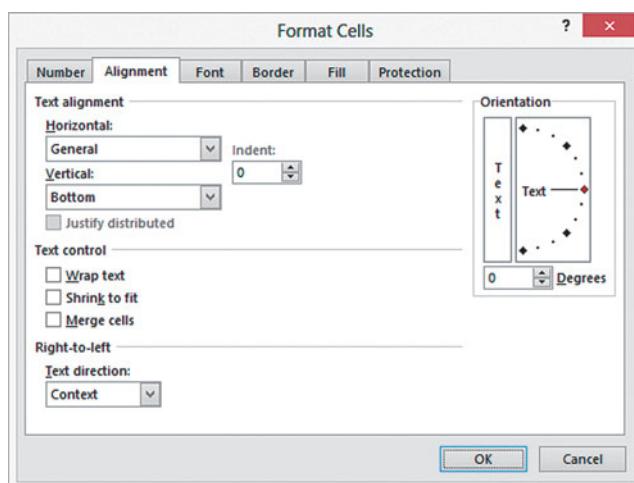
Clicking a **Category** changes the panel to the right of the list. For example, clicking **Number** displays a panel (partially shown below) in which you can set the number of decimal places. (Many cells in the worksheets used in this book have been set to display four decimal places.)



You can also change the numeric formatting of cells by clicking the various buttons of the **Number** group in the Home tab (shown below).



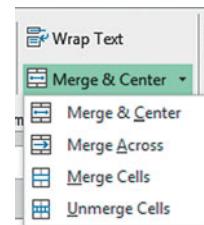
When you click the **Alignment** tab of the Format Cells dialog box (shown below), you display a panel in which you can control such things such as whether cell contents get displayed centered or top- or bottom-anchored in a cell and whether cell contents are horizontally centered or left or right justified.



These choices in this panel are duplicated in the Alignment group of the Home tab (shown below).



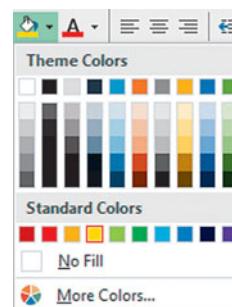
In the Ribbon interface, many buttons, such as **Merge & Center**, have an associated drop-down list that you display by clicking the drop-down arrow at the right. For Merge & Center, this drop-down displays a gallery of similar choices (shown below).



While the **Font** tab of the Format Cells dialog box allows you to change the text attributes used to display cell contents, consider using the equivalent choices in the **Font** group of the Home tab (shown below). Using this group is a more convenient way of making choices such as changing the typeface or point size or styling text to be bold or italic.

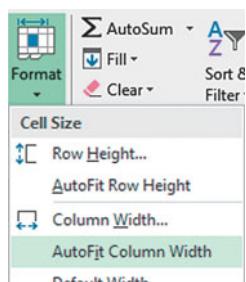


To change the background color of a cell, click the **fill** icon in the Font group. Clicking this icon changes the background color to the color that appears behind the bucket (yellow in the illustration below). Clicking the drop-down button to the right of the fill icon displays a gallery of colors (shown below) from which you can select a color or click **More Colors** for even more choices. (The letter A icon and its drop-down button offer similar choices for the color of the text being displayed.)



To adjust the width of a column to an optimal size, select the column and then select **Format → Autofit Column Width** (shown on page 726) in the Cells group. Excel will

adjust the width of the column to accommodate the display of the values in all of the cells of the column.



B.6 Chart Formatting

Microsoft Excel does not always use best practices when constructing charts. Many of the *In-Depth Excel* instructions that involve charts refer you to this section so that you can correct the formatting of a chart that was just constructed. To apply any of the following corrections, you must first select the chart that is to be corrected. (If Chart Tools or PivotChart Tools appears above the Ribbon tabs, you have selected a chart.)

If, when you open to a chart sheet, the chart is either too large to be fully seen or too small and surrounded by a frame mat that is too large, click **Zoom Out** or **Zoom In**, located in the lower-right portion of the Excel window frame, to adjust the chart display.

In the following, instructions preceded with **(2013)** apply only to Excel 2013. Unlike other Excel versions, in Excel 2013 some of the selections, such as the gridlines selections, are toggles that turn on (or off) a chart element.

Changes You Most Commonly Make

To relocate a chart to its own chart sheet:

1. Click the chart background and click **Move Chart** from the shortcut menu.
2. In the Move Chart dialog box, click **New Sheet**, enter a name for the new chart sheet, and click **OK**.

To turn off the improper horizontal gridlines:

(2013) **Design** → **Add Chart Element** → **Gridlines** → **Primary Major Horizontal**

Layout → **Gridlines** → **Primary Horizontal**
Gridlines → **None**

To turn off the improper vertical gridlines:

(2013) **Design** → **Add Chart Element** → **Gridlines** → **Primary Major Horizontal**

Layout → **Gridlines** → **Primary Vertical**
Gridlines → **None**

To turn off the chart legend:

(2013) **Design** → **Add Chart Element** → **Legend** → **None**

Layout → **Legend** → **None**

If you use Excel 2007, you will also need to apply these changes:

Layout → **Data Labels** → **None**

Layout → **Data Table** → **None**

These two apply *only* to Excel 2007.

Chart and Axis Titles

To add a chart title to a chart missing a title:

1. In Excel 2013, select **Design** → **Add Chart Element** → **Chart Title** → **Above Chart**. Otherwise, click on the chart and then select **Layout** → **Chart Title** → **Above Chart**.
2. In the box that is added to the chart, select the words “Chart Title” and enter an appropriate title.

To add a title to a horizontal axis missing a title:

1. In Excel 2013, select **Design** → **Add Chart Element** → **Axis Titles** → **Primary Horizontal**. Otherwise, click on the chart and then select **Layout** → **Axis Titles** → **Primary Horizontal Axis Title** → **Title Below Axis**.
2. In the box that is added to the chart, select the words “Axis Title” and enter an appropriate title.

To add a title to a vertical axis missing a title:

1. In Excel 2013, select **Design** → **Add Chart Element** → **Axis Titles** → **Primary Vertical**. Otherwise, click on the chart and then select **Layout** → **Axis Titles** → **Primary Vertical Axis Title** → **Rotated Title**.
2. In the box that is added to the chart, select the words “Axis Title” and enter an appropriate title.

Chart Axes

To turn on the display of the X axis, if not already shown:

(2013) **Design** → **Add Chart Element** → **Axes** → **Primary Horizontal**

Layout → **Axes** → **Primary Horizontal Axis** → **Show Left to Right Axis** (or **Show Default Axis**, if listed)

To turn on the display of the Y axis, if not already shown:

(2013) **Design** → **Add Chart Element** → **Axes** → **Primary Vertical**

Layout → **Axes** → **Primary Vertical Axis** → **Show Default Axis**

For a chart that contains secondary axes, to turn off the secondary horizontal axis title:

(2013) **Design** → **Add Chart Element** → **Axis Titles** → **Secondary Horizontal**

Layout → **Axis Titles** → **Secondary Horizontal** → **Axis Title** → **None**

For a chart that contains secondary axes, to turn on the secondary vertical axis title:

(2013) Design → Add Chart Element → Axis Titles → Secondary Vertical

Layout → Axis Titles → Secondary Vertical Axis Title → Rotated Title

Correcting the Display of the X Axis

In scatter plots and related line charts, Microsoft Excel displays the *X* axis at the *Y* axis origin ($Y = 0$). For plots that have negative values, this causes the *X* axis not to appear at the bottom of the chart. To relocate the *X* axis so that it appears at the bottom of a scatter plot or line chart, open to the chart sheet containing the chart and:

1. Right-click the ***Y* axis** and click **Format Axis** from the shortcut menu.

In the Format Axis dialog box:

2. Click **Axis Options** in the left pane. In the Axis Options pane on the right, click **Axis value** and in its box enter the value shown in the dimmed **Minimum** box (near the top of the pane).
3. Click **Close**.

Emphasizing Histogram Bars

To better emphasize each bar in a histogram, open to the chart sheet containing the histogram and:

1. Right-click over one of the histogram bars and click **Format Data Series** in the shortcut menu.

In the Format Data Series dialog box:

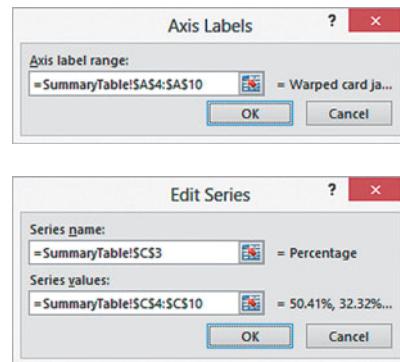
2. Click **Border Color** in the left pane. In the Border Color right pane, click **Solid line**. From the Color dropdown list, click the darkest color in the same column as the currently selected (highlighted) color.
3. Click **Border Styles** in the left pane. In the Border Styles right pane, click the up spinner button to set the **Width** to **3 pt**.
4. Click **OK**.

B.7 Selecting Cell Ranges for Charts

Selecting Cell Ranges for Chart Labels and Series

As a general rule, you either type that cell range or select the cell range by using the mouse pointer to enter a cell range in a Microsoft Excel dialog box. You are free to choose to enter the cell range either using relative or absolute references (see Section B.2). The Axis Labels and Edit Series dialog boxes, associated with chart labels and data series, are two

exceptions. These dialog boxes and their contents for the Pareto chart sheet of the Pareto workbook are shown below.



To enter a cell range into these two dialog boxes, you must enter the cell range as a *formula* that uses absolute cell references in the form *WorksheetName!UpperLeftCell:LowerRightCell*. This is best done using the mouse-pointer method to enter these cell ranges. Typing the cell range, as you might normally do, will often be frustrating, as keys such as the cursor keys will not function as they do in other dialog boxes.

Selecting a Non-contiguous Cell Range

Typically, you enter a non-contiguous cell range such as the cells A1:A11 and C1:C11 by typing the cell range of each group of cells, separated by commas—for example, **A1:A11, C1:C11**. In certain contexts, such as using the dialog boxes discussed in the preceding section, you will need to select that non-contiguous cell range using the mouse pointer method. To use the mouse-pointer method with such ranges, first, select the cell range of the first group of cells and then, while holding down **Ctrl**, select the cell range of the other groups of cells that form the non-contiguous cell range.

B.8 Deleting the “Extra” Bar from a Histogram

As explained in “Classes and Excel Bins” on page 45, you use bins to approximate classes. One result of this approximation is that you will always create an “extra” bin that will have a frequency of zero. Because, by definition, this extra bin considers values that are less than the lowest value that exists in your set of data and therefore will always have the frequency zero, you can safely and properly eliminate the “extra” bar that represents this bin.

To do so, you need to edit the cell range that Excel uses to construct the histogram. Right-click the histogram background and click **Select Data**. In the Select Data Source Data dialog box, first click **Edit** under the **Legend Entries (Series)** heading. In the Edit Series dialog box, edit the **Series values** cell range formula to begin with the second cell of the

original cell range and click **OK**. Then click **Edit** under the **Horizontal (Categories) Axis Labels** heading. In the Axis Labels dialog box, edit the **Axis label range** to begin with the second cell of the original cell range and click **OK**.

B.9 Creating Histograms for Discrete Probability Distributions

You can create a histogram for a discrete probability distribution based on a discrete probabilities table. For example, to create the Figure 5.3 histogram of the binomial probability distribution on page 199, open to the **COMPUTE worksheet** of the **Binomial workbook**. Select the cell range **B14:B18**, the probabilities in the Binomial Probabilities Table, and:

1. Select **Insert → Column** and select the first **2-D Column** gallery choice (**Clustered Column**).
2. Right-click the chart background and click **Select Data**.

In the Select Data Source dialog box:

3. Click **Edit** under the **Horizontal (Categories) Axis Labels** heading.
4. In the Axis Labels dialog box, enter the cell range $formula = \text{COMPUTE!A14:A18}$ as the **Axis label range**. (See Section B.7 to learn how to best enter this cell range formula.) Click **OK** to return to the Select Data Source dialog box.
5. Back in the Select Data Source dialog box, click **OK**.

In the chart:

6. Right-click inside a bar and click **Format Data Series** in the shortcut menu.

In the Format Data Series dialog box:

7. Click **Series Options** in the left pane. In the Series Options right pane, change the **Gap Width** slider to **Large Gap**. Click **Close**.

Relocate the chart to a chart sheet and adjust the chart formatting by using the instructions in Section B.6.

C.1 About the Online Resources for This Book

Online resources support your study of business statistics and your use of this book. Online resources are available from a special download web page for this book as well as in a MyStatLab course for this book. On the download page, these resources are packaged as a series of zip archive files, one zip file for each of the categories listed below. In the MyStatLab course for this book, online resources are also available on a chapter-by-chapter basis. Categories of online resources are:

- **Excel and Minitab Data files** The files that contain the data used in chapter examples, named in problems, or used in the end-of-chapter cases, including the *Managing Ashland MultiComm Services* running case. A complete list of these files and their contents appears in Section C.3.
- **Excel Guide Workbooks** Excel workbooks that contain templates or model solutions for applying Excel to a particular statistical method. A complete list of the Excel Guide Workbooks appear in *Excel Guide Workbooks* in Section C.3.
- **Files for the Digital Cases** The set of PDF files that support the end-of-chapter Digital Cases. Some of the Digital Case PDF files contain attached or embedded Excel data workbooks for use with particular case questions.
- **Online Topics** The set of PDF format files that present additional statistical topics. Included in this set is the full text of two chapters, “Statistical Applications in Quality Management” and “Decision Making.”
- **Short Takes** The set of PDF files that extend the discussion of specific concepts or further document the results presented in the book.
- **Visual Explorations Workbooks** The workbooks that interactively demonstrate various key statistical concepts. Three of these workbooks are add-in workbooks stored in the **.xlam** Excel add-in format. See *Visual Explorations* in Section C.3 for additional information.

As explained in Section EG.1 on page 8, this book also supports the use of PHStat, the Pearson Education statistics add-in for Microsoft Excel. If you plan to use PHStat with this book, see Section C.4.

Accessing the Online Resources

Online resources for this book are available either on the student download page for this book or inside the MyStatLab course for this book (see Section C.3). To access resources from the student download page for this book:

1. Visit www.pearsonhighered.com/levine.
2. In that web page, find the entries for this book, *Basic Business Statistics*, thirteenth edition, and click the student download page link.
3. In the download page, click the link for the desired items. Most items will cause the web browser to prompt you to save the (zip archive) that you can save and later unzip. Some download links may require an access code (see Section C.2).

C.2 Accessing the MyStatLab Course Online

The MyStatLab course for this book contains all the online resources for this book. Log into the course and in the left panel of the course page for this book, click **Tools for Success**. On that page, click the link for one of the online resource categories listed in Section C.1. Selected files and data workbooks may also be available in the chapter-by-chapter resource pages.

Using MyStatLab requires an access code. An access code may have been packaged with this book. If your book did not come with an access code, you can obtain one at mypearson.com.

C.3 Details of Downloadable Files

Data Files

Data workbooks contain the data used in chapter examples or named in problems. Throughout this book, the names of data workbooks appear in a special inverted color typeface—for example, **Retirement Funds**. Data files are stored as worksheets in both the **.xlsx** Excel workbook and the **.mtw** Minitab worksheet file formats. (For files that contain more than one worksheet, Minitab versions are stored as **.mpj** Minitab project files.) Worksheets organize the data for each variable by column, using the rules discussed in Sections EG.2 and MG.2.

In the following alphabetical list, the variables for each data file are presented in the order of their appearance, starting with first column (A in Excel and C1 in Minitab). Chapter references indicate the chapter or chapters that use the data file in an example or problem. A trailing (E) notes a file exclusive to Excel. A trailing (M) notes a file exclusive to Minitab.

311CALLCENTER Day and abandonment rate (%) (Chapter 3)

ACCOUNTINGPARTNERS Firm and number of partners (Chapter 3)

ACCOUNTINGPARTNERS2 Region and number of partners (Chapter 10)

ACCOUNTINGPARTNERS4 Region and number of partners (Chapter 11)

ACCOUNTINGPARTNERS6 Region, revenue (\$millions), number of partners, number of professionals, MAS (%), southeast (0 = no, 1 = yes) Gulf coast southeast (0 = no, 1 = yes) (Chapter 15)

ACT Method (online or traditional), ACT scores for condensed course, ACT scores for regular course (Chapter 11)

ACT-ONEWAY Group 1 ACT scores, group 2 ACT scores, group 3 ACT scores, group 4 ACT scores (Chapter 11)

ADINDEX Respondent, cola A Adindex, and cola B Adindex (Chapter 10)

ADVERTISE Sales (\$thousands), radio ads (\$thousands), and newspaper ads (\$thousands) for 22 cities (Chapters 14, 15, and 17)

ADVERTISING Sales (\$millions) and newspaper ads (\$thousands) (Chapter 15)

AMS2-1 Types of errors and frequency, types of errors and cost, types of wrong billing errors and cost (as three separate worksheets) (Chapter 2)

AMS2-2 Days and number of calls (Chapter 2)

AMS8 Rate willing to pay (\$) (Chapter 8)

AMS9 Upload speed (Chapter 9)

AMS10 Update times for email interface 1 and email interface 2 (Chapter 10)

AMS11-1 Update time for system 1, system 2, and system 3 (Chapter 11)

AMS11-2 Technology (cable or fiber) and interface (system 1, system 2, or system 3) (Chapter 11)

AMS13 Number of hours spent telemarketing and number of new subscriptions (Chapter 13)

AMS14 Week, number of new subscriptions, hours spent telemarketing, and type of presentation (formal or informal) (Chapter 14)

AMS16 Month and number of home delivery subscriptions (Chapter 16)

ANSCOMBE Data sets A, B, C, and D, each with 11 pairs of X and Y values (Chapter 13)

ATM TRANSACTIONS Cause, frequency, and percentage (Chapter 2)

AUDITS Year and number of audits (Chapters 2 and 16)

AUTOMAKER1 Automaker and number of complaints (Chapters 2 and 17)

AUTOMAKER2 Category and number of complaints (Chapter 2)

AUTOSALES Manufacturer, sales, and change percentage (Chapter 17)

BANK1 Waiting time (in minutes) of 15 customers at a bank located in a commercial district (Chapters 3, 9, 10, and 12)

BANK2 Waiting time (in minutes) of 15 customers at a bank located in a residential area (Chapters 3, 10, and 12)

BANKMARKETING Age, type of job, marital status (divorced, married, or single), education (primary, secondary, tertiary, or unknown), is credit in default, mean yearly balance in account, is there a housing loan, is there a personal loan, last contact duration in seconds, number of contacts performed during this campaign, has the client purchased a term deposit (also contains the BinaryLogisticDATA worksheet that contains recoded variables) (Chapter 18)

BASEBALL Team; attendance; high temperature on game day; winning percentage of home team; opponent's winning percentage; game played on weekend day (0 = no, 1 = yes) and promotion held (0 = no, 1 = yes) (Chapter 18)

BB2012 Team, league (0 = American, 1 = National) wins, earned run average, runs scored, hits allowed, walks allowed, saves, and errors (Chapters 13, 14, 15, and 17)

BBCOST2012 Team and fan cost index (Chapters 2, 6, and 17)

BBREVENUE2013 Team, revenue (\$millions), and value (\$millions) (Chapter 13)

BBSALARIES Year and average major league baseball salary (\$millions) (Chapter 16)

BED & BATH Year, coded year, and number of stores open (Chapter 16)

BESTFUND\$1 Fund type (short-term or long-term), 1-year return, and 3-year return (Chapters 10 and 17)

BESTFUND\$2 Fund type (short-term, long-term, or world), 1-year return, and 3-year return (Chapter 11)

BESTFUND\$3 Fund type (small, mid-cap, or large), 1-year return, and 3-year return (Chapter 11)

BOOKPRICES Author, title, bookstore price, and online price (\$) (Chapter 10)	1-to-5 ordinal scale (1 = poor to 5 = excellent), annual household income (\$), and average number of miles the customer expects to walk/run each week (Chapters 2, 3, 6, 8, 10, 11, and 12)
BRAKES Part, gauge 1, and gauge 2 (Chapter 11)	CARDSTUDY Upgraded (0 = no, 1 = yes), purchases (\$thousands), and extra cards (0 = no, 1 = yes) (Chapters 14 and 17)
BRANDZTECHFIN Brand, brand value in 2011 (\$millions), % change in brand value from 2010, region, and sector (Chapters 10 and 12)	CARPRODUCTION Year, coded year, and number of units produced (Chapter 16)
BRANDZTECHFINTELE Brand, brand value in 2011 (\$millions), % change in brand value from 2010, region, and sector (Chapter 11)	CATFOOD Ounces eaten of kidney, shrimp, chicken liver, salmon, and beef cat food (Chapters 11 and 12)
BREAKFAST Type (Continental or American), delivery time difference for early time period, and delivery time difference for late time period (Chapter 11)	CATFOOD2 Piece size (F = fine, C = chunky), coded weight for low fill height, and coded weight for current fill height (Chapter 11)
BREAKFAST2 Type (Continental or American), delivery time difference for early time period, and delivery time difference for late time period (Chapter 11)	CDRATE Bank, 1-year CD rate, and 5-year CD rate (Chapters 2, 3, 6, and 8)
BRYNNEPACKAGING WPCT score and rating (Chapter 13)	CEO-COMPENSATION Company, CEO compensation (\$millions), and return in 2012 (Chapters 2, 3, and 13)
BULBS Manufacturer (1 = A, 2 = B) and length of life (hours) (Chapters 2, 10, and 12)	CEREALS Cereal, calories, carbohydrates, and sugar (Chapters 3, 13, and 17)
BUNDLE Restaurant, bundle score, and typical cost (\$) (Chapter 2)	CHALLENGING Data and charts for Figure 2.19 (Chapter 2)
BUSINESSVALUATION Drug company name, price to book value ratio, return on equity (ROE), and growth% (Chapters 14 and 17)	CHURN Customer ID, churn coded (0 = No, 1 = Yes), churn, calls, and visits (Chapters 14 and 17)
BUSINESSVALUATION2 Company, ticker symbol, Standard Industrial Classification 3 (SIC3) code, Standard Industrial Classification 4 (SIC4) code, price to book value ratio, price to earnings ratio, natural log of assets (as a measure of size), return on equity (ROE), growth percentage (GS5), debt to EBITDA ratio, dummy variable indicator of SIC 4 code 2834 (1 = 2834, 0 = not 2824), and dummy variable indicator of SIC 4 code 2835 (1 = 2835, 0 = not 2835) (Chapter 15)	CIGARETTETAX State and cigarette tax (\$) (Chapters 2 and 3)
CABERNET California wine rating, Washington wine rating, California wine ranking, and Washington wine ranking (Chapter 12)	CIRCUITS Batch, position 1, position 2, position 18, position 19, and position 28 (Chapter 11)
CAFFEINE Caffeine per fluid ounce (mg/oz) (Chapter 2)	COCA-COLA Year, coded year, and revenues (\$billions) (Chapter 16)
CALLCENTER Month and call volume (Chapter 16)	COFFEE Expert and rating of coffees by brand A, B, C, and D (Chapters 10 and 11)
CAMERAS Subcompact battery life, compact battery life (Chapters 10 and 12)	COFFEEPRICESPORTUGAL Year and retail price of coffee in Portugal (€/kg) (Chapter 16)
CANDIDATE ASSESSMENT Salary, competence rating, gender of candidate (F or M), gender of rater (F or M), rater/candidate gender (F to F, F to M, M to M, M to F), school (Private, Public), department (Biology, Chemistry, Physics), and age of rater (Chapter 18)	COFFEESALES Coffee sales at \$0.59, \$0.69, \$0.79, and \$0.89 (Chapters 11 and 12)
CARDIOGOODFITNESS Product purchased (TM195, TM498, TM798), age in years, gender (Male or Female), education in years, relationship status (Single or Partnered), average number of times the customer plans to use the treadmill each week, self-rated fitness on a	COFFEESALES2 Coffee sales and price (Chapter 15)
	COLA Beverage end-cap sales and produce end-cap sales (Chapters 10 and 12)
	COLLEGE FOOTBALL Head coach, school, conference, school pay of head coach, other pay, total pay, max bonus, and football net revenue (Chapters 2, 3, and 13)
	COMPUTERSALES Year, coded year, and computer and software sales (\$millions) (Chapter 16)
	CONCRETE1 Sample number and compressive strength after two days and seven days (Chapter 10)
	CONCRETE2 Sample number and compressive strength after 2, 7, and 28 days (Chapter 11)

CONGESTION City, annual time waiting in traffic (hours), and cost of waiting in traffic (\$) (Chapters 2 and 3)	FALSEIMPRESSIONS Data and charts for selected Section 2.7 figures (Chapter 2) (E)
COSTESTIMATION Units produced and total cost (\$) (Chapter 15)	FASTFOOD Amount spent on fast food (\$) (Chapters 2, 3, 8, and 9)
CPI-U Year, coded year, and value of CPI-U (the consumer price index) (Chapter 16)	FEDRECEIPT Year, coded year, and federal receipts (\$billions current) (Chapter 16)
CREDIT SCORES City, state, and average credit score (Chapters 2 and 3)	FIFTEENWEEKS Week number, number of customers, and sales (\$thousands) over a period of 15 consecutive weeks (Chapter 13)
CURRENCY Year, coded year, and exchange rates (against the U.S. dollar) for the Canadian dollar, Japanese yen, and English pound sterling (Chapters 2 and 16)	FIVEYEARCDRATE Five-year CD rates in New York and Los Angeles (Chapter 10)
CURRENCY2 Currency and value of US\$1 for years from 2002 through 2012 (Chapter 17)	FLYASH Fly ash percentage and strength (PSI) (Chapter 15)
DELIVERY Customer, number of cases, and delivery time (Chapter 13)	FOODS Type, bland/spicy rating, light/heavy rating, low/high calories rating (Chapter 17)
DENSITY Ammonium percentage, density for stir rate of 100, and density for stir rate of 150 (Chapter 11)	FORCE Force required to break an insulator (Chapters 2, 3, 8, and 9)
DOINGBUSINESS Region, country name, 2012 GDP per capita, Internet users 2011 (per 100 people), and mobile cellular subscriptions 2011 (per 100 people) (Chapter 17)	FOREIGNMARKET Country, level of development (Emerging or Developed), and time required to start a business (days) (Chapter 10)
DOMESTICBEER Brand, alcohol percentage, calories, and carbohydrates (Chapters 2, 3, 6, and 15)	FOREIGNMARKET2 Country, region, cost to export container (US\$), cost to import container (US\$) (Chapters 11 and 12)
DOMESTICBEER2 Brand, calories, and carbohydrates (Chapter 17)	FREPORT Address, fair market value (\$thousands), property size (acres), house size, age, number of rooms, number of bathrooms, and number of cars that can be parked in the garage (Chapter 15)
DOWDOGS Stock and 1-year return (Chapter 3)	FRESHFOOD Fresh food, United States pounds per capita consumed, Japan pounds per capita consumed, and Russia pounds per capita consumed (Chapter 2)
DOWMARKETCAP Company and market capitalization (\$billions) (Chapters 3 and 6)	FTGLOBAL500 Sector (Automobiles & parts, Financial services, Health care equipment & services, or Software & computer services), country, company, market cap (\$billions), and 52-week change (%) (Chapter 17)
DRILL Depth, time to drill additional 5 feet, and type of hole (dry or wet) (Chapters 14 and 17)	FTMBA School name, program cost (\$), total students per program, job offers (%), and mean starting salary (\$) (Chapters 15 and 17)
DRINK Amount of soft drink filled in 2-liter bottles (Chapters 2 and 9)	FURNITURE Days between receipt and resolution of complaints regarding purchased furniture (Chapters 2, 3, 8, and 9)
DRIVE-THRUSPEED Year and drive-thru speed in seconds (Chapter 16)	GASPRICES Month and price per gallon (\$) (Chapter 16)
ENERGY State and per capita kilowatt hour use (Chapter 3)	GCFREEROSLYN Address, location (Glen Cove, Freeport, or Roslyn), fair market value (\$thousands), property size (acres), age, house size (sq. ft.), number of rooms, number of bathrooms, and number of cars that can be parked in the garage (Chapter 15)
ENTREE Type and number served (Chapter 2)	GCROSLYN Address, location (Glen Cove or Roslyn), fair market value (\$thousands), property size (acres), age, house size (sq. ft.), number of rooms, number of bathrooms, and number of cars that can be parked in the garage (Chapters 14 and 15)
ERWAITING Emergency room waiting time (in minutes) at the main facility and at satellite 1, satellite 2, and satellite 3 (Chapters 11 and 12)	GDP Year and gross domestic product (Chapter 16)
ESPRESSO Tamp (inches) and time (seconds) (Chapter 13)	
EUROTOURISM Country, employment in tourism 2012, tourism establishments (Chapter 15)	
EUROTOURISM2 Country, employment in tourism 2012, business travel & tourism spending 2012 (US\$millions), international visitors 2012, tourism establishments (Chapter 18)	
FACEBOOKTIME Gender (F or M) and amount of time in minutes spent on Facebook per day (Chapter 9)	
FACEBOOKTIME2 Gender (F or M) and amount of time in minutes spent on Facebook per day (Chapter 10)	

GLENCOVE Address, fair market value (\$thousands), property size (acres), age, house size (sq. ft.), number of rooms, number of bathrooms, and number of cars that can be parked in the garage (Chapters 14, 15, and 17)

GLOBALSOCIALMEDIA Country, GDP, and social media usage (%) (Chapters 2, 3, 13, and 17)

GOLD Quarter, coded quarter, price (\$), Q1, Q2, and Q3 (Chapter 16)

GOLFBALL Distance for designs 1, 2, 3, and 4 (Chapters 11 and 12)

GPIGMAT GMAT scores and GPA (Chapter 13)

GRADSURVEY ID, gender (Female or Male), age (as of last birthday), graduate major (Accounting, CIS, Economics/Finance, International Business, Management, Retailing/Marketing, or Other), current graduate GPA, undergraduate major (Biological Sciences, Business, Engineering, or Other), undergraduate GPA, current employment status (Full-Time, Part-Time, or Unemployed), number of different full-time jobs held in the past 10 years, expected salary upon completion of MBA (\$thousands), amount spent for books and supplies this semester (\$), advisory rating, type of computer owned (Desktop or Laptop), text messages per week, wealth accumulated to feel rich (Chapters 2, 3, 6, 8, 10, 11, and 12)

GRANULE Granule loss in Boston and Vermont shingles (Chapters 3, 8, 9, and 10)

HEATINGOIL Monthly consumption of heating oil (gallons), temperature (degrees Fahrenheit), attic insulation (inches), ranch-style (0 = not ranch-style, 1 = ranch-style) (Chapters 14 and 15)

HEMLOCKF FARMS Asking price, hot tub (0 = no, 1 = yes), rooms, lake view (0 = no, 1 = yes), bathrooms, bedrooms, loft/den (0 = no, 1 = yes), finished basement (0 = no, 1 = yes), number of acres (Chapter 15)

HOTELAWAY Nationality and cost (British pounds sterling) (Chapter 3)

HOTELPRICES City and average price (US\$) of a hotel room at a 2-star price, 3-star price, and 4-star hotel (Chapters 2 and 3)

HOUSEHOLDS ID, Gender, age, Hispanic origin (N or Y), dwelling type (AB, AH, DH, or Other), age of dwelling (years), years living at dwelling, number of bedrooms, number of vehicles kept at dwelling, fuel type at dwelling (Electric, Gas, Oil, or Other), monthly cost of fuel at dwelling (\$), U.S. citizenship (N or Y), college degree (N or Y), marital status (D, M, NM, S, or W), work for pay in previous week (N or Y), mode of transportation to work (Bus, Car, Home, Subway/Rail, Taxi, Other, or NA), commuting time (minutes), hours worked per week, type of organization(GOV, NA, PP, PNP, or SE), annual earned income (\$), and total annual income (\$) (Chapter 18)

HYBRIDSALES Year and number sold (Chapter 18)

ICECREAM Daily temperature (in degrees Fahrenheit) and sales (\$thousands) for 21 days (Chapter 13)

INDICES Year, change in DJIA, S&P500, and NASDAQ (Chapter 3)

INSURANCE Processing time in days for insurance policies (Chapters 3, 8, and 9)

INSURANCECLAIMS Claims, buildup (0 = buildup not indicated, 1 = buildup indicated), excess payment (\$) (Chapter 8)

INSURANCEFRAUD ID, fraud coded (0 = no, 1 = yes), fraud (No or Yes), new business coded (0 = no, 1 = yes), new business (No or Yes), and claims/year (Chapters 14 and 17)

INVOICE Number of invoices processed and amount of time (hours) for 30 days (Chapter 13)

INVOICES Amount recorded (in dollars) from sales invoices (Chapter 9)

LUGGAGE Delivery time (in minutes) for luggage in Wing A and Wing B of a hotel (Chapters 10 and 12)

MANAGERS Sales (ratio of yearly sales divided by the target sales value for that region), Wonderlic Personnel Test score, Strong-Campbell Interest Inventory Test score, number of years of selling experience prior to becoming a sales manager, whether the sales manager has a degree in electrical engineering (No or Yes) (Chapter 15)

MARKET PENETRATION Country and Facebook penetration (in percentage) (Chapters 3 and 8)

MCDONALDS Year, coded year, and annual total revenues (\$billions) at McDonald's Corporation (Chapter 16)

MEDICALWIRES1 Machine type, narrow, and wide (Chapter 11)

MEDICALWIRES2 Narrow, and wide (Chapter 11)

METAL INDICES Metal and the total rate of return (%) for the years 2006 through 2012 (Chapter 17)

METALS Year and the total rate of return (in percentage) for platinum, gold, and silver (Chapter 3)

MINING Day, amount stacked, and downtime (Chapter 18)

MINING2 Day; hours of downtime due to mechanical, electrical, tonnage restriction, operator, and no feed; and total hours (Chapter 18)

MMCDRATE Bank, and interest rates for money market, 1-year CD, 2-year CD, and 5-yr CD (Chapter 11)

MOBILE ELECTRONICS STACKED Stacked version of Mobile Electronics (Chapter 11) (M)

MOBILE ELECTRONICS In-aisle sales, front sales, kiosk sales, and expert area sales (Chapter 11) (E)

MOBILE ELECTRONICS2 Mobile payments (No or Yes), in-aisle sales, front sales, kiosk sales, and expert area sales (Chapter 11)

- MOISTURE** Moisture content of Boston shingles and Vermont shingles (Chapter 9)
- MOLDING** Vibration time (seconds), vibration pressure (psi), vibration amplitude (%), raw material density (g/mL), quantity of raw material (scoops), product length in cavity 1 (in.), product length in cavity 2 (in.), product weight in cavity 1 (gr.), and product weight in cavity 2 (gr.) (Chapter 15)
- MOTIVATION** Factor, mean rating by global employees, and mean rating by U.S. employees (Chapter 10)
- MOVIE** Title, box office gross (\$millions), and DVD revenue (\$millions) (Chapter 13)
- MOVIE ATTENDANCE** Year and movie attendance (billions) (Chapters 2 and 16)
- MOVIE ATTENDANCE2** Movie attendance (billions) for years 2002 through 2012 (Chapter 17)
- MOVIE REVENUES** Year and revenue (\$billions) (Chapter 2)
- MOVING** Labor hours, cubic feet, number of large pieces of furniture, and availability of an elevator (Chapters 13, 14, and 17)
- MYELOMA** Patient, before transplant measurement, after transplant measurement (Chapter 10)
- NATURAL GAS** Month, wellhead price (\$/thousands cu. ft.), and residential price (\$/thousands cu. ft.) (Chapter 2)
- NATURAL GAS2** Month, wellhead price (\$/thousands cu. ft.), and residential price (\$/thousands cu. ft.) (Chapter 16)
- NBA2012** Team, number of wins, points scored, points allowed, point difference, field goal %, field goal % allowed, field goal % difference, own turnovers, opponent turnovers, turnover difference, and rebounds % (Chapters 14 and 18)
- NBAVALUES** Team, team code, annual revenue (\$millions), and value (\$millions) and 1-year change in value (%) (Chapters 2, 3, 13, and 17)
- NEEDS** Need and frequency (Chapter 2)
- NEIGHBOR** Selling price (\$thousands), number of rooms, neighborhood location (0 = east, 1 = west) (Chapter 14)
- NEWHOMESALES** Month, sales in thousands, and mean price (\$thousands) (Chapter 2)
- OIL&GASOLINE** Week, price of a gallon of gasoline (\$), and price of oil per barrel, (\$) (Chapter 13)
- OMNIPOWER** Bars sold, price (cents), and promotion expenses (\$) (Chapters 14 and 17)
- ONLINE SHOPPING** Main reason and percentage (Chapter 2)
- ORDER** Time in minutes to fill orders for a population of 200 (Chapter 8)
- ORGANICFOOD** Customer, organic food purchaser (0 = no, 1 = yes), age, online health wellness e-newsletters subscriber (0 = no, 1 = yes) (Chapter 14)
- O-RING** Flight number, temperature, and O-ring damage index (Chapter 13)
- PACKAGEDFOOD** Category, United States sales, Japan sales, Russia sales (Chapter 17)
- PACKAGINGFOAM1** Die temperature, 3 mm. diameter, and 4 mm. diameter (Chapter 11)
- PACKAGINGFOAM2** Die temperature, 3 mm. diameter, and 4 mm. diameter (Chapter 11)
- PACKAGINGFOAM3** Die temperature, die diameter, and foam density (Chapter 14)
- PACKAGINGFOAM4** Die temperature, die diameter, and foam diameter (Chapter 14)
- PAINRELIEF** Temperature and dissolve times for Equate, Kroger, and Alka-Seltzer tablets (Chapter 11)
- PALLET** Weight of Boston shingles and weight of Vermont shingles (Chapters 2, 8, 9, and 10)
- PARACHUTE1WAY** Tensile strength of parachutes from suppliers 1, 2, 3, and 4 (Chapter 11)
- PARACHUTE2WAY** Loom and tensile strength of parachutes from suppliers 1, 2, 3, and 4 (Chapter 11)
- PEN** Ad and product rating (Chapters 11, 12)
- PHILLY** Zip code, population, median sales price 2012 (\$000), average days on market 2012, units sold 2012, median household income (\$), percentage of residents with a BA or higher, and hotness (0 = not hot, 1 = hot) (Chapter 18)
- PHONE** Time (in minutes) to clear telephone line problems and location (1 = I, 2 = II) (Chapters 10 and 12)
- PIZZAHUT** Gender coded (0 = Female, 1 = Male), gender (Female or Male), price (\$), and purchase (0 = student selected another pizzeria, 1 = student selected Pizza Hut) (Chapters 14 and 17)
- PIZZATIME** Time period, delivery time for local restaurant, and delivery time for national chain (Chapter 10)
- POLIO** Year and incidence rates per 100,000 persons of reported poliomyelitis (Chapter 16)
- POTATO** Percentage of solids content in filter cake, acidity (pH), lower pressure, upper pressure, cake thickness, varidrive speed, and drum speed setting for 54 measurements (Chapter 15)
- POTTERMOVIES** Title, first weekend gross (\$millions), U.S. gross (\$millions), and worldwide gross (\$millions) (Chapters 2, 3, 13, and 17)
- PROPERTYTAXES** State and property taxes per capita (\$) (Chapters 2 and 3)
- PROTEIN** Type of food, calories (in grams), protein, percentage of calories from fat, percentage of calories from saturated fat, and cholesterol (mg) (Chapters 2, 3 and 17)
- PUMPKIN** Circumference and weight of pumpkins (Chapter 13)

- QSR** Company, average sales per unit, market segment (Chapters 11 and 12)
- QSRCHAIN** Evaluator and ratings for Henry St, Surf Av, Granby, and Blvd N restaurants (Chapter 11)
- REDANDWHITE** Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, wine type coded (0 = White, 1 = Red), wine type (Red or White), quality (Chapter 14)
- REDWOOD** Height (ft.), breast height diameter (in.), and bark thickness (in.) (Chapters 13 and 14)
- REGISTRATIONERROR** Registration error, temperature, pressure, and supplier (Chapter 15)
- REGISTRATIONERROR-HIGHCOST** Registration error and temperature (Chapter 15)
- RENTSILVERSPRING** Apartment size (sq. ft.) and monthly rental cost (\$) (Chapter 13)
- RESTAURANTS** Location (City or Suburban), food rating, decor rating, service rating, summated rating, coded location (0 = City, 1 = Suburban), and cost of a meal (Chapters 2, 3, 10, 13, and 14)
- RESTAURANTS2** Location, food rating, decor rating, service rating, cost of a meal, popularity index, and cuisine [American (New), Chinese, French, Indian, Italian, Japanese, or Mexican] (Chapter 18)
- RETIREMENT FUNDS** Fund number, market cap (Small, Mid-Cap, or Large), type (Growth or Value), assets (\$millions), turnover ratio, beta (measure of the volatility of a stock), standard deviation (measure of returns relative to 36-month average), risk (Low, Average, or High), 1-year return, 3-year return, 5-year return, 10-year return, expense ratio, star rating (Chapters 2, 3, 6, 8, 10, 11, 12, 15, and 17)
- ROSLYN** Address, fair market value (\$thousands), property size (acres), house size (sq. ft.), age, number of rooms, number of bathrooms, and number of cars that can be parked in the garage (Chapter 15)
- SATISFACTION** Satisfaction code (0 = not satisfied, 1 = satisfied), Satisfaction (No or Yes), delivery time difference (minutes), previous coded (0 = no, 1 = yes), and previous (No or Yes) (Chapter 17)
- SECOND EXPERIMENT** In-aisle sales, front sales, kiosk sales, and expert area sales (Chapter 12) (E)
- SECOND EXPERIMENT STACKED** Stacked version of Second Experiment (Chapter 12) (M)
- SEDANS** Miles per gallon for 2013 midsized sedans (Chapters 3 and 8)
- SERVICETIME** Fast food chain and mean service time for years 1 through 12 (Chapter 17)
- SILVER** Year and price of silver (\$) (Chapter 16)
- SILVER-Q** Quarter, coded quarter, price of silver (\$), Q1, Q2, and Q3 (Chapter 16)
- SILVERSPRING** Address, asking price (\$000), assessed value (\$000), taxes (\$), size (thousands sq. ft.) fireplace coded (0 = no, 1 = yes), number of bedrooms, number of bathrooms, age (years), fireplace (No or Yes) (Chapters 13, 14, and 15)
- SITESELECTION** Store number, profiled customers, and sales (\$millions) (Chapter 13)
- SMARTPHONES** Price (\$) (Chapter 3)
- SMARTPHONES SALES** Type, and market share percentage for the years 2011 through 2013 (Chapter 2)
- SOCCERVALUES2013** Team, country, revenue (\$millions), and value (\$millions) (Chapter 13)
- SOLAR POWER** Year and amount of solar power generated (megawatts) (Chapter 16)
- SPILLS** Year and number of oil spills in the Gulf of Mexico (Chapter 16)
- SPORTS** Sport, movement rating, speed rating, rules rating, team oriented rating, and amount of contact rating (Chapter 17)
- STANDBY** Standby hours, total staff present, remote hours, Dubner hours, and total labor hours (Chapters 14 and 15)
- STARBUCKS** Tear, viscosity, pressure, plate gap (Chapters 13, 14 and 17)
- STEEL** Error in actual length and specified length (Chapters 2, 6, 8, and 9)
- STOCK INDICES** Stock index and percentage change for 2006 through 2012 (Chapter 17)
- STOCK PERFORMANCE** Decade and stock performance (%) (Chapters 2 and 16)
- STOCKPRICES2012** Date, S&P 500 value, and closing weekly stock price for GE, Discovery Communications, and Google (Chapter 13)
- STUDYTIME** Gender and study time in hours (Chapter 10)
- SUPERMARKETPRICES** Shopping item and price at Publix, Winn-Dixie, Target, and Walmart (Chapter 11)
- SUV** Miles per gallon for 2013 small SUVs (Chapters 3, 6, and 8)
- TABLE_5.1 X and P(X)** (Chapter 5) (M)
- TARGETWALMART** Shopping item, Target price (\$), and Walmart price (\$) (Chapter 10)
- TAX** Quarterly sales tax receipts (\$thousands) (Chapter 3)
- TEABAGS** Weight of tea bags in ounces (Chapters 3, 8, and 9)
- TELECOM** Provider, TV rating, and Phone rating (Chapter 10)
- TELECOM2** Provider, TV rating, Phone rating, and Internet rating (Chapter 11)
- TESTRANK** Rank scores and training method (0 = traditional, 1 = experimental) for 10 people (Chapter 12)
- THICKNESS** Thickness, catalyst, pH, pressure, temperature, and voltage (Chapters 14 and 15)

- THREE-HOTEL SURVEY** Choose again? (No or Yes) and Golden Palm, Palm Royale, and Palm Princess tallies (Chapter 12) (M)
- TIMES** Get-ready times (Chapter 3)
- TOYS R US** Quarter, coded quarter, revenue, and the dummy variables Q1, Q2, and for quarters (Chapter 16)
- TRAVEL** Month and amount of travel (Chapter 16)
- TROUGH** Width of trough (Chapters 2, 3, 8, and 9)
- TRSNYC** Year, unit value of Diversified Equity funds, and unit value of Stable Value funds (Chapter 16)
- TSMODEL1** Year, coded year, and three time series (I, II, and III) (Chapter 16)
- TSMODEL2** Year, coded year, and two time series (I and II) (Chapter 16)
- TWITTERMOVIES** Movie, Twitter activity, and receipts (\$) (Chapter 13)
- TWO-HOTEL SURVEY** Choose again? (No or Yes) and Beachcomber and Windsurfer tallies (Chapter 12) (M)
- UNDERGRADSURVEY** ID, gender (Female or Male), age (as of last birthday), class designation (Sophomore, Junior, or Senior), major (Accounting, CIS, Economics/Finance, International Business, Management, Retail/Marketing, Other, or Undecided), graduate school intention (No, Yes, or Undecided), cumulative GPA, current employment status (Full-Time, Part-Time, or Unemployed), expected starting salary (\$thousands), number of social networking sites registered for, satisfaction with student advisement services on campus, amount spent on books and supplies this semester, type of computer preferred (Desktop, Laptop, or Tablet), text messages per week, wealth accumulated to feel rich (Chapters 2, 3, 6, 8, 10, 11, and 12)
- UNDERWRITING** End-of-training exam score, proficiency exam score, and training method (classroom, courseware app, or online) (Chapter 14)
- UNSTACKED 1YRRETURN** One-year return percentage for growth funds and one-year return percentage for value funds (Chapter 2) (M)
- USED CARS** Car, year, age, price (\$), mileage, power (hp), fuel (mpg) (Chapter 18)
- UTILITY** Utilities charges (\$) for 50 one-bedroom apartments (Chapters 2 and 6)
- VB** Time to complete program (Chapter 10)
- VINHOVERDE** Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality (Chapters 13, 14 and 15)
- VINHOVERDE POPULATION** Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, wine type (Red or White), and quality (Chapter 17)
- WAIT** Waiting time and seating time (Chapter 6)
- WALMART** Quarter and Wal-Mart Stores quarterly revenues (\$billions) (Chapter 16)
- WARECOST** Distribution cost (\$thousands), sales (\$thousands), and number of orders (Chapters 13, 14, and 15)
- WIP** Processing times at each of two plants (1 = A, 2 = B) (Chapter 18)
- WL_SOCIALDATA** Land, ride, comments, rating, and alternate scale (Chapter 17)
- WL_WAITDATA** Land, ride, and wait (minutes) (Chapter 17)
- WL_WAITHISTORY** Ride and historical wait times (minutes) at 21 half-hour intervals (Chapter 17)
- WORKFORCE** Year, population, and size of the workforce (Chapter 16)
- YARN** Side-by-side aspect and breaking strength scores for 30 psi, 40 psi, and 50 psi (Chapter 11)

Excel Guide Workbooks

Excel Guide workbooks contain templates or model solutions for applying Excel to a particular statistical method. Chapter examples and the *In-Depth Excel* instructions of the Excel Guides feature worksheets from these workbooks and PHStat constructs many of the worksheets from these workbooks for you.

Workbooks are stored in the .xlsx Excel workbook format. Most contain a **COMPUTE worksheet** (often shown in this book) that presents results as well as a **COMPUTE_FORMULAS worksheet** that allows you to examine all of the formulas used in the worksheet. The Excel Guide workbooks (with the number of the chapter in which each is first mentioned) are:

Recoded (1)	Sample Size Mean (8)
Random (1)	Sample Size Proportion (8)
Data Cleaning (1)	Z Mean workbook (9)
Summary Table (2)	T mean workbook (9)
Contingency Table (2)	Z Proportion (9)
Distributions (2)	Pooled-Variance T (10)
Pareto (2)	Separate-Variance T (10)
Stem-and-leaf (2)	Paired T (10)
Histogram (2)	F Two Variances (10)
Polygons (2)	Z Two Proportions (10)
Scatter Plot (2)	One-Way ANOVA (11)
Time Series (2)	Levene (11)
MCT (2)	Randomized Block (11)
Descriptive(3)	Chi-Square (12)
Quartiles (3)	Chi-Square Worksheets (12)
Boxplot (3)	Wilcoxon (12)
Parameters(3)	Kruskal-Wallis Worksheets (12)
VE-Variability (3)	Simple Linear Regression (13)
Covariance (3)	Package Delivery (13)
Probabilities (4)	Multiple Regression (14)
Bayes (4)	Logistic Regression add-in (14)
Discrete Variable (5)	Moving Averages (16)
Portfolio (5)	Exponential Smoothing (16)
Binomial (5)	Exponential Trend (16)
Poisson (5)	Differences (16)
Hypergeometric (5)	Lagged Predictors (16)
Normal (6)	Forecasting Comparison (16)
NPP (6)	GaugeBullet (17)
Exponential (6)	Treemap (17)
SDS (7)	Slicers (17)
CIE sigma known (8)	
CIE sigma unknown (8)	
CIE Proportion (8)	

The **Slicers workbook** works only with Microsoft Windows versions Excel 2010 and Excel 2013. The **Logistic Regression add-in workbook** requires the Solver add-in and, if you use a Microsoft Windows version of Excel, the security settings discussed in Appendix Section D.3. (Appendix Section D.6 discusses how to check for the presence of the Solver add-in.)

Visual Explorations

Visual Explorations are workbooks that interactively demonstrate various key statistical concepts. Three workbooks are add-in workbooks that are stored in the .xlam Excel add-in format. Using these add-in workbooks with Microsoft Windows Excel requires the security settings discussed in Appendix Section D.3. The visual exploration workbooks are:

- VE-Normal Distribution (add-in)
- VE-Sampling Distribution (add-in)
- VE-Simple Linear Regression (add-in)
- VE-Variability

PDF Files

PDF files use the Portable Document Format that can be viewed in most web browsers or with PDF utility programs, such as Adobe Reader, a free program available at get.adobe.com/reader/. Both the Digital Case files and the online sections use this format.

C.4 PHStat

PHStat is the Pearson Education statistics add-in for Microsoft Excel that simplifies the task of using Excel as you learn business statistics. PHStat comes packaged as a zip file archive that you download and unzip to the folder of your choice. The archive contains:

PHStat.xlam, the actual add-in workbook that is further discussed in Appendix Sections D.2 and G.1, and these four supporting files:

PHStat readme.pdf Explains the technical requirements, and setup and troubleshooting procedures for PHStat (PDF format).

PHStatHelp.chm The integrated help system for users of Microsoft Windows Excel.

PHStatHelp.pdf The help system as a PDF format file.

PHStatHelp.epub The help system for eBook reader applications (Open Publication Structure eBook format).

Downloading PHStat requires an access code. If your book was packaged with an access code, then click the PHStat link on either the student download page or the MyStatLab **Tools for Success** page (see Sections C.1 and C.2) to be taken to the PHStat home page. On that page, click the download link and follow the instructions for entering the access code.

If your book was not packaged with an access code, visit the PHStat home page (www.pearsonhighered.com/phstat) to learn how you can obtain a code.

APPENDIX D Configuring Microsoft Excel

This appendix seeks to eliminate the common types of technical problems that could complicate your use of Microsoft Excel as you learn business statistics with this book. You will want to be familiar with the contents of this appendix—and follow all its directives—if the copy of Microsoft Excel you plan to use runs on a computer system that you control and maintain. If you use a computer system that is maintained by others, such as a computer system in an academic computer lab, this appendix can be a useful resource for those in charge of solving technical issues that may arise.

Not all sections of this appendix apply to all readers. Sections with the code (WIN) apply to you if you use Microsoft Excel with Microsoft Windows, while sections with the code (OS X) apply to you if you use Microsoft Excel with OS X (formerly, Mac OS X). Some sections apply to all readers (ALL). (If you use Minitab, there are no configuration issues that you need to address.)

D.1 Getting Microsoft Excel Ready for Use (ALL)

You must have an up-to-date, properly licensed copy of Microsoft Excel in order to work through the examples and solve the problems in this book as well as to take advantage of the Excel-related workbooks and add-ins described in Appendix C. To get Microsoft Excel ready for use, follow this checklist:

- If necessary, install Microsoft Excel on your computer system.
- Check and apply Microsoft-supplied updates to Microsoft Excel and Microsoft Office.
- After you first use Microsoft Excel, recheck for Microsoft-supplied updates at least once every two weeks.

If you need to install a new copy of Microsoft Excel on a Microsoft Windows computer system, choose the 32-bit version and not the 64-bit version *even if you have a 64-bit version of a Microsoft Windows operating system*. Many people mistakenly believe that the 64-bit version is somehow “better,” not realizing that the OS X Excel 2011 is a 32-bit version and that Microsoft advises you to choose the 32-bit version for reasons the company details on its website. (The 64-bit WIN version exists primarily for users who need to work with Excel workbooks that are greater than 2GB in size. What would a 2GB workbook store? By one informal calculation, the contents of over 60 copies of this book—in other words, *big data*, as defined in Section GS.3.)

Checking For and Applying Updates

Microsoft Excel updates require Internet access and the process to check for and apply updates differs among Excel versions. If you use a Microsoft Windows version of Excel and use Windows 7 or 8, checking for updates is done by the Windows Update service. If you use an older version of Microsoft Windows, you may have to upgrade to this service (Visit the Microsoft Download Center at <http://www.microsoft.com/download/default.aspx> for further details.)

Windows Update can automatically apply any updates it finds, although many users prefer to set Windows Update to *notify* when updates are available and then select and apply updates manually.

In OS X Excel versions and some Microsoft Windows versions, you can manually check for updates. In Excel 2011 (OS X), select **Help → Check for Updates** and in the dialog box that appears, click **Check for Updates**. In Excel 2007 (WIN), first click the **Office Button** and then **Excel Options** at the bottom of the Office Button window. In the Excel options dialog box, click **Resources** in the left pane and then in the right pane click **Check for Updates** and follow the instructions that appear on the web page that is displayed.

You normally do not manually check for updates in either Excel 2010 (WIN) or Excel 2013 (WIN). However, in some installations of these versions, you can select **File □ Account □ Update Options** (2013) or **File □ Help □ Check for Updates** (2010) and select options or follow instructions to manually check for updates.

If all else fails, you can open a web browser and go to the Microsoft Office part of the Microsoft Download Center at www.microsoft.com/download/office.aspx?q=office and manually select and download updates. On the web page that gets displayed, filter the downloadable files by specifying the Excel version you use. Discover the version number and update status by these means:

- In Excel 2013 (WIN), select **File □ Account** and then click **About Microsoft Excel**. In the dialog box that appears note the numbers and codes that follow the phrase “Microsoft Excel 2013.”
- In Excel 2010 (WIN), select **File □ Help**. Under the heading “About Microsoft Excel” click **Additional Version and Copyright Information** and in the dialog box that appears note the numbers and codes that follow “Microsoft Excel 2010.”
- In Excel 2011 (OS X), click **Excel □ About Excel**. The dialog box that appears displays the **Version** and **Latest Installed Update**.
- In Excel 2007 (WIN), first click the **Office Button** and then click **Excel Options**. In the Excel options dialog box, click **Resources** in the left pane. In the right pane note the numbers and codes that follow Microsoft Office Excel 2007 under the “about Microsoft Office Excel 2007” heading.

Special Note for Office 365 Users

If you use Office 365, you are using the most current version of Excel for your system. At the time of publication, the most current version for Microsoft Windows systems was Excel 2013. For OS X, the most current version was Excel 2011.

D.2 Getting PHStat Ready for Use (ALL)

If you plan to use PHStat, the Pearson Education add-in workbook that simplifies the use of Microsoft Excel with this book (see Section EG.1 on page 8), you must first download PHStat using an access code as discussed in Section C.4. The PHStat download comes packaged as a zip file archive that you unzip to the folder of your choice.

PHStat is fully compatible with these Excel versions: Excel 2007 (WIN), Excel 2010 (WIN), Excel 2011 (OS X), and Excel 2013 (WIN). PHStat is not compatible with Excel 2008 (OS X), an Excel version that did not include the capability of running add-in workbooks. If you are using Microsoft Excel with Microsoft Windows (any version), then you must first configure the Microsoft

Excel security settings as discussed in Section D.3. If you are using Microsoft Excel with OS X, no additional steps are required.

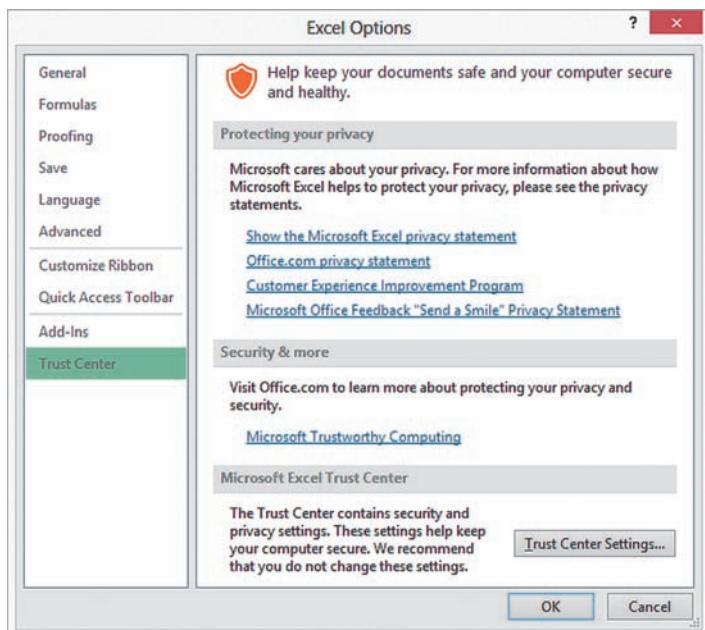
D.3 Configuring Excel Security for Add-In Usage (WIN)

The Microsoft Excel security settings can prevent add-ins such as PHStat and the Visual Explorations add-in workbooks from opening or functioning properly. To configure these security settings to permit proper PHStat functioning:

1. In Excel 2010 and Excel 2013, select **File □ Options**. In Excel 2007, first click the **Office Button** and then click **Excel Options**.

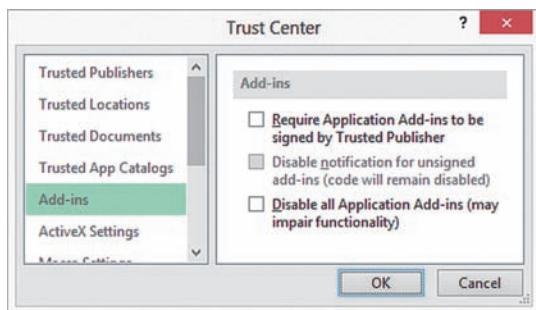
In the Excel Options dialog box (shown below):

2. Click **Trust Center** in the left pane and then click **Trust Center Settings** in the right pane.

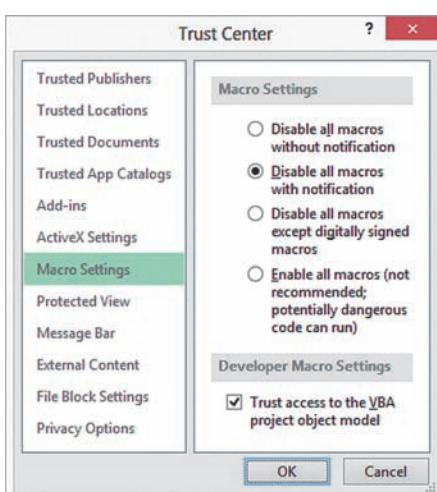


In the Trust Center dialog box:

3. Click **Add-ins** in the next left pane, and in the Add-ins right pane clear all of the checkboxes (shown below).



4. Click **Macro Settings** in the left pane, and in the Macro Settings right pane click **Disable all macros with notification** and check **Trust access to the VBA object model** (shown below).



5. Click **OK** to close the Trust Center dialog box.

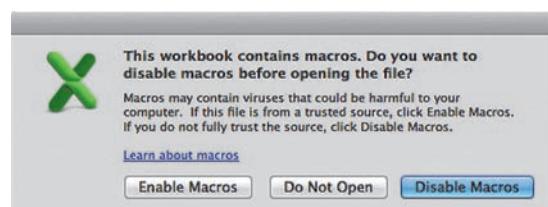
Back in the Excel Options dialog box:

6. Click **OK** to finish.

On some systems that have stringent security settings, you might need to modify step 4. For such systems, in step 4, also click **Trusted Locations** in the left pane and then, in the Trusted Locations right pane, click **Add new location** to add the folder path that you chose to store the PHStat files.

D.4 Opening PHStat (ALL)

Open the **PHStat.xlam** file to use PHStat. As you open the file, by any of the means discussed in Section EG.3 on page 9, Microsoft Excel will display a warning dialog box. The dialog boxes for Excel 2013 (WIN) and Excel 2011 (OS X) are shown below. Click **Enable Macros**, which is not the default choice, to enable PHStat to function properly.



After you click **Enable Macros**, you can verify that PHStat has opened properly by looking for a PHStat menu in the Add-Ins tab of the Office Ribbon (WIN) or in the menu at top of the display (OS X).

If you have skipped checking for and applying necessary Excel updates, or if some of the updates were unable to be applied, when you first attempt to use PHStat, you may see a “Compile Error” message that talks about a “hidden module.” If this occurs, repeat the process of checking for and applying updates to Excel. Review the PHStat FAQs in Appendix G for additional assistance, if necessary, and occasionally check for PHStat updates by revisiting the page from which you originally downloaded the PHStat files.

D.5 Using a Visual Explorations Add-in Workbook (ALL)

To use any of the Visual Explorations add-in workbooks, you must first download them using one of the methods discussed in Appendix Section C.1. If your download is packaged as a zip archive file, you must unzip that archive and store the add-in workbook files together in a folder of your choosing. Then apply the Section D.3 instructions, if necessary. When you open a Visual Explorations add-in workbook, you will see the same type of warning dialog box that Section D.4. describes. Click **Enable Macros** to enable the workbook to function properly.

D.6 Checking for the Presence of the Analysis ToolPak or Solver Add-Ins (ALL)

If you choose to perform logistic regression using the Section EG14.7 *PHStat* or *In-Depth Excel* instructions, you will need to ensure that the Solver add-in has been installed. Similarly, if you choose to use the *Analysis ToolPak* Excel Guide instructions, you will need to ensure that the

Microsoft Excel Analysis ToolPak add-in has been installed. (This add-in is not available if you use Microsoft Excel with OS X.)

To check for the presence of the Solver (or Analysis ToolPak) add-in, if you use Microsoft Excel with Microsoft Windows:

1. Select **File → Options**. (In Excel 2007, click the **Office Button** and then click **Excel Options**.)

In the Excel Options dialog box:

2. Click **Add-Ins** in the left pane and look for the entry **Solver Add-in** (or **Analysis ToolPak**) in the right pane, under **Active Application Add-ins**.
3. If the entry appears, click **OK**.
4. If the entry does not appear in the **Active Application Add-ins** list, select **Excel Add-ins** from the **Manage** drop-down list and then click **Go**.
5. In the Add-Ins dialog box, check **Solver Add-in** (or **Analysis ToolPak**) in the **Add-Ins available** list and click **OK**.

If Analysis ToolPak (or Solver Add-in) does not appear in the list, rerun the Microsoft Office setup program to install this component.

To check for the presence of the Solver add-in, if you use Microsoft Excel with OS X, select **Tools → Options**. In the Add-Ins dialog box, check **Solver.Xlam** in the **Add-Ins available** list and click **OK**.

APPENDIX E Tables

TABLE E.1

Table of Random Numbers

Row	Column							
	00000 12345	00001 67890	11111 12345	11112 67890	22222 12345	22223 67890	33333 12345	33334 67890
01	49280	88924	35779	00283	81163	07275	89863	02348
02	61870	41657	07468	08612	98083	97349	20775	45091
03	43898	65923	25078	86129	78496	97653	91550	08078
04	62993	93912	30454	84598	56095	20664	12872	64647
05	33850	58555	51438	85507	71865	79488	76783	31708
06	97340	03364	88472	04334	63919	36394	11095	92470
07	70543	29776	10087	10072	55980	64688	68239	20461
08	89382	93809	00796	95945	34101	81277	66090	88872
09	37818	72142	67140	50785	22380	16703	53362	44940
10	60430	22834	14130	96593	23298	56203	92671	15925
11	82975	66158	84731	19436	55790	69229	28661	13675
12	30987	71938	40355	54324	08401	26299	49420	59208
13	55700	24586	93247	32596	11865	63397	44251	43189
14	14756	23997	78643	75912	83832	32768	18928	57070
15	32166	53251	70654	92827	63491	04233	33825	69662
16	23236	73751	31888	81718	06546	83246	47651	04877
17	45794	26926	15130	82455	78305	55058	52551	47182
18	09893	20505	14225	68514	47427	56788	96297	78822
19	54382	74598	91499	14523	68479	27686	46162	83554
20	94750	89923	37089	20048	80336	94598	26940	36858
21	70297	34135	53140	33340	42050	82341	44104	82949
22	85157	47954	32979	26575	57600	40881	12250	73742
23	11100	02340	12860	74697	96644	89439	28707	25815
24	36871	50775	30592	57143	17381	68856	25853	35041
25	23913	48357	63308	16090	51690	54607	72407	55538
26	79348	36085	27973	65157	07456	22255	25626	57054
27	92074	54641	53673	54421	18130	60103	69593	49464
28	06873	21440	75593	41373	49502	17972	82578	16364
29	12478	37622	99659	31065	83613	69889	58869	29571
30	57175	55564	65411	42547	70457	03426	72937	83792
31	91616	11075	80103	07831	59309	13276	26710	73000
32	78025	73539	14621	39044	47450	03197	12787	47709
33	27587	67228	80145	10175	12822	86687	65530	49325
34	16690	20427	04251	64477	73709	73945	92396	68263
35	70183	58065	65489	31833	82093	16747	10386	59293
36	90730	35385	15679	99742	50866	78028	75573	67257
37	10934	93242	13431	24590	02770	48582	00906	58595
38	82462	30166	79613	47416	13389	80268	05085	96666
39	27463	10433	07606	16285	93699	60912	94532	95632
40	02979	52997	09079	92709	90110	47506	53693	49892
41	46888	69929	75233	52507	32097	37594	10067	67327
42	53638	83161	08289	12639	08141	12640	28437	09268
43	82433	61427	17239	89160	19666	08814	37841	12847
44	35766	31672	50082	22795	66948	65581	84393	15890
45	10853	42581	08792	13257	61973	24450	52351	16602
46	20341	27398	72906	63955	17276	10646	74692	48438
47	54458	90542	77563	51839	52901	53355	83281	19177
48	26337	66530	16687	35179	46560	00123	44546	79896
49	34314	23729	85264	05575	96855	23820	11091	79821
50	28603	10708	68933	34189	92166	15181	66628	58599

TABLE E.1

Table of Random
Numbers (continued)

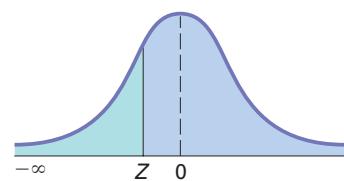
Row	Column							
	00000 12345	00001 67890	11111 12345	11112 67890	22222 12345	22223 67890	33333 12345	33334 67890
51	66194	28926	99547	16625	45515	67953	12108	57846
52	78240	43195	24837	32511	70880	22070	52622	61881
53	00833	88000	67299	68215	11274	55624	32991	17436
54	12111	86683	61270	58036	64192	90611	15145	01748
55	47189	99951	05755	03834	43782	90599	40282	51417
56	76396	72486	62423	27618	84184	78922	73561	52818
57	46409	17469	32483	09083	76175	19985	26309	91536
58	74626	22111	87286	46772	42243	68046	44250	42439
59	34450	81974	93723	49023	58432	67083	36876	93391
60	36327	72135	33005	28701	34710	49359	50693	89311
61	74185	77536	84825	09934	99103	09325	67389	45869
62	12296	41623	62873	37943	25584	09609	63360	47270
63	90822	60280	88925	99610	42772	60561	76873	04117
64	72121	79152	96591	90305	10189	79778	68016	13747
65	95268	41377	25684	08151	61816	58555	54305	86189
66	92603	09091	75884	93424	72586	88903	30061	14457
67	18813	90291	05275	01223	79607	95426	34900	09778
68	38840	26903	28624	67157	51986	42865	14508	49315
69	05959	33836	53758	16562	41081	38012	41230	20528
70	85141	21155	99212	32685	51403	31926	69813	58781
71	75047	59643	31074	38172	03718	32119	69506	67143
72	30752	95260	68032	62871	58781	34143	68790	69766
73	22986	82575	42187	62295	84295	30634	66562	31442
74	99439	86692	90348	66036	48399	73451	26698	39437
75	20389	93029	11881	71685	65452	89047	63669	02656
76	39249	05173	68256	36359	20250	68686	05947	09335
77	96777	33605	29481	20063	09398	01843	35139	61344
78	04860	32918	10798	50492	52655	33359	94713	28393
79	41613	42375	00403	03656	77580	87772	86877	57085
80	17930	00794	53836	53692	67135	98102	61912	11246
81	24649	31845	25736	75231	83808	98917	93829	99430
82	79899	34061	54308	59358	56462	58166	97302	86828
83	76801	49594	81002	30397	52728	15101	72070	33706
84	36239	63636	38140	65731	39788	06872	38971	53363
85	07392	64449	17886	63632	53995	17574	22247	62607
86	67133	04181	33874	98835	67453	59734	76381	63455
87	77759	31504	32832	70861	15152	29733	75371	39174
88	85992	72268	42920	20810	29361	51423	90306	73574
89	79553	75952	54116	65553	47139	60579	09165	85490
90	41101	17336	48951	53674	17880	45260	08575	49321
91	36191	17095	32123	91576	84221	78902	82010	30847
92	62329	63898	23268	74283	26091	68409	69704	82267
93	14751	13151	93115	01437	56945	89661	67680	79790
94	48462	59278	44185	29616	76537	19589	83139	28454
95	29435	88105	59651	44391	74588	55114	80834	85686
96	28340	29285	12965	14821	80425	16602	44653	70467
97	02167	58940	27149	80242	10587	79786	34959	75339
98	17864	00991	39557	54981	23588	81914	37609	13128
99	79675	80605	60059	35862	00254	36546	21545	78179
100	72335	82037	92003	34100	29879	46613	89720	13274

Source: Partially extracted from the Rand Corporation, *A Million Random Digits with 100,000 Normal Deviates* (Glencoe, IL, The Free Press, 1955).

TABLE E.2

The Cumulative Standardized Normal Distribution

Entry represents area under the cumulative standardized normal distribution from $-\infty$ to Z



Z	Cumulative Probabilities									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-6.0	0.000000001									
-5.5	0.000000019									
-5.0	0.000000287									
-4.5	0.000003398									
-4.0	0.000031671									
-3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003
-3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
-3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
-3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
-3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
-3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
-3.3	0.00048	0.00047	0.00045	0.00043	0.00042	0.00040	0.00039	0.00038	0.00036	0.00035
-3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
-3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
-3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00103	0.00100
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2388	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2482	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

TABLE E.2

The Cumulative Standardized Normal Distribution (continued)

Entry represents area under the cumulative standardized normal distribution from $-\infty$ to Z

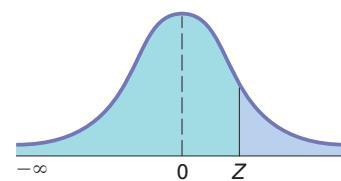
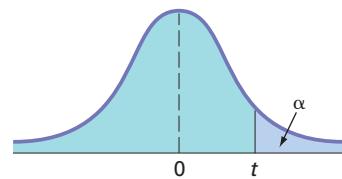


TABLE E.3Critical Values of t

For a particular number of degrees of freedom, entry represents the critical value of t corresponding to the cumulative probability $(1 - \alpha)$ and a specified upper-tail area (α).



Degrees of Freedom	Cumulative Probabilities					
	0.75	0.90	0.95	0.975	0.99	0.995
	Upper-Tail Areas					
0.25	0.10	0.05	0.025	0.01	0.005	
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5177	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500
31	0.6825	1.3095	1.6955	2.0395	2.4528	2.7440
32	0.6822	1.3086	1.6939	2.0369	2.4487	2.7385
33	0.6820	1.3077	1.6924	2.0345	2.4448	2.7333
34	0.6818	1.3070	1.6909	2.0322	2.4411	2.7284
35	0.6816	1.3062	1.6896	2.0301	2.4377	2.7238
36	0.6814	1.3055	1.6883	2.0281	2.4345	2.7195
37	0.6812	1.3049	1.6871	2.0262	2.4314	2.7154
38	0.6810	1.3042	1.6860	2.0244	2.4286	2.7116
39	0.6808	1.3036	1.6849	2.0227	2.4258	2.7079
40	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045
41	0.6805	1.3025	1.6829	2.0195	2.4208	2.7012
42	0.6804	1.3020	1.6820	2.0181	2.4185	2.6981
43	0.6802	1.3016	1.6811	2.0167	2.4163	2.6951
44	0.6801	1.3011	1.6802	2.0154	2.4141	2.6923
45	0.6800	1.3006	1.6794	2.0141	2.4121	2.6896
46	0.6799	1.3002	1.6787	2.0129	2.4102	2.6870
47	0.6797	1.2998	1.6779	2.0117	2.4083	2.6846
48	0.6796	1.2994	1.6772	2.0106	2.4066	2.6822
49	0.6795	1.2991	1.6766	2.0096	2.4049	2.6800
50	0.6794	1.2987	1.6759	2.0086	2.4033	2.6778

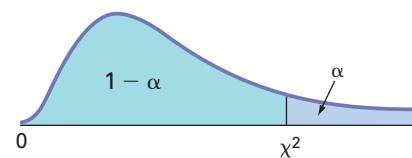
TABLE E.3Critical Values of t (continued)

For a particular number of degrees of freedom, entry represents the critical value of t corresponding to the cumulative probability $(1 - \alpha)$ and a specified upper-tail area (α) .

Degrees of Freedom	Cumulative Probabilities					
	0.75	0.90	0.95	0.975	0.99	0.995
	Upper-Tail Areas					
0.25	0.10	0.05	0.025	0.01	0.005	
51	0.6793	1.2984	1.6753	2.0076	2.4017	2.6757
52	0.6792	1.2980	1.6747	2.0066	2.4002	2.6737
53	0.6791	1.2977	1.6741	2.0057	2.3988	2.6718
54	0.6791	1.2974	1.6736	2.0049	2.3974	2.6700
55	0.6790	1.2971	1.6730	2.0040	2.3961	2.6682
56	0.6789	1.2969	1.6725	2.0032	2.3948	2.6665
57	0.6788	1.2966	1.6720	2.0025	2.3936	2.6649
58	0.6787	1.2963	1.6716	2.0017	2.3924	2.6633
59	0.6787	1.2961	1.6711	2.0010	2.3912	2.6618
60	0.6786	1.2958	1.6706	2.0003	2.3901	2.6603
61	0.6785	1.2956	1.6702	1.9996	2.3890	2.6589
62	0.6785	1.2954	1.6698	1.9990	2.3880	2.6575
63	0.6784	1.2951	1.6694	1.9983	2.3870	2.6561
64	0.6783	1.2949	1.6690	1.9977	2.3860	2.6549
65	0.6783	1.2947	1.6686	1.9971	2.3851	2.6536
66	0.6782	1.2945	1.6683	1.9966	2.3842	2.6524
67	0.6782	1.2943	1.6679	1.9960	2.3833	2.6512
68	0.6781	1.2941	1.6676	1.9955	2.3824	2.6501
69	0.6781	1.2939	1.6672	1.9949	2.3816	2.6490
70	0.6780	1.2938	1.6669	1.9944	2.3808	2.6479
71	0.6780	1.2936	1.6666	1.9939	2.3800	2.6469
72	0.6779	1.2934	1.6663	1.9935	2.3793	2.6459
73	0.6779	1.2933	1.6660	1.9930	2.3785	2.6449
74	0.6778	1.2931	1.6657	1.9925	2.3778	2.6439
75	0.6778	1.2929	1.6654	1.9921	2.3771	2.6430
76	0.6777	1.2928	1.6652	1.9917	2.3764	2.6421
77	0.6777	1.2926	1.6649	1.9913	2.3758	2.6412
78	0.6776	1.2925	1.6646	1.9908	2.3751	2.6403
79	0.6776	1.2924	1.6644	1.9905	2.3745	2.6395
80	0.6776	1.2922	1.6641	1.9901	2.3739	2.6387
81	0.6775	1.2921	1.6639	1.9897	2.3733	2.6379
82	0.6775	1.2920	1.6636	1.9893	2.3727	2.6371
83	0.6775	1.2918	1.6634	1.9890	2.3721	2.6364
84	0.6774	1.2917	1.6632	1.9886	2.3716	2.6356
85	0.6774	1.2916	1.6630	1.9883	2.3710	2.6349
86	0.6774	1.2915	1.6628	1.9879	2.3705	2.6342
87	0.6773	1.2914	1.6626	1.9876	2.3700	2.6335
88	0.6773	1.2912	1.6624	1.9873	2.3695	2.6329
89	0.6773	1.2911	1.6622	1.9870	2.3690	2.6322
90	0.6772	1.2910	1.6620	1.9867	2.3685	2.6316
91	0.6772	1.2909	1.6618	1.9864	2.3680	2.6309
92	0.6772	1.2908	1.6616	1.9861	2.3676	2.6303
93	0.6771	1.2907	1.6614	1.9858	2.3671	2.6297
94	0.6771	1.2906	1.6612	1.9855	2.3667	2.6291
95	0.6771	1.2905	1.6611	1.9853	2.3662	2.6286
96	0.6771	1.2904	1.6609	1.9850	2.3658	2.6280
97	0.6770	1.2903	1.6607	1.9847	2.3654	2.6275
98	0.6770	1.2902	1.6606	1.9845	2.3650	2.6269
99	0.6770	1.2902	1.6604	1.9842	2.3646	2.6264
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259
110	0.6767	1.2893	1.6588	1.9818	2.3607	2.6213
120	0.6765	1.2886	1.6577	1.9799	2.3578	2.6174
∞	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758

TABLE E.4Critical Values of χ^2

For a particular number of degrees of freedom, entry represents the critical value of χ^2 corresponding to the cumulative probability $(1 - \alpha)$ and a specified upper-tail area (α).

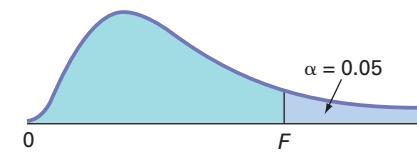


Degrees of Freedom	Cumulative Probabilities											
	0.005	0.01	0.025	0.05	0.10	0.25	0.75	0.90	0.95	0.975	0.99	0.995
	0.995	0.99	0.975	0.95	0.90	0.75	0.25	0.10	0.05	0.025	0.01	0.005
1			0.001	0.004	0.016	0.102	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.455	7.841	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	4.255	9.037	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	5.071	10.219	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	5.899	11.389	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	6.737	12.549	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	7.584	13.701	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	8.438	14.845	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	9.299	15.984	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	10.165	17.117	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	11.037	18.245	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	11.912	19.369	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	12.792	20.489	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	13.675	21.605	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	14.562	22.718	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	15.452	23.828	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	16.344	24.935	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	17.240	26.039	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	18.137	27.141	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	19.037	28.241	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	19.939	29.339	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	20.843	30.435	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	21.749	31.528	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	22.657	32.620	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	23.567	33.711	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	24.478	34.800	40.256	43.773	46.979	50.892	53.672

For larger values of degrees of freedom (df) the expression $Z = \sqrt{2\chi^2} - \sqrt{2(df - 1)}$ may be used and the resulting upper-tail area can be found from the cumulative standardized normal distribution (Table E.2).

TABLE E.5Critical Values of F

For a particular combination of numerator and denominator degrees of freedom, entry represents the critical values of F corresponding to the cumulative probability $(1 - \alpha)$ and a specified upper-tail area (α) .

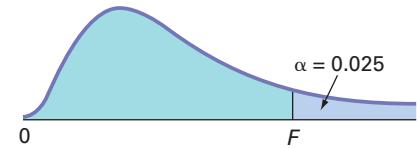


Cumulative Probabilities = 0.95																			
		Upper-Tail Areas = 0.05																	
		Numerator, df_1																	
Denominator, df_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90	243.90	245.90	248.00	249.10	250.10	251.10	252.20	253.30	254.30
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	1.91	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	1.67	1.57	1.52	1.46	1.39	1.32	1.22

(continued)

TABLE E.5Critical Values of F (continued)

For a particular combination of numerator and denominator degrees of freedom, entry represents the critical values of F corresponding to the cumulative probability $(1 - \alpha)$ and a specified upper-tail area (α) .

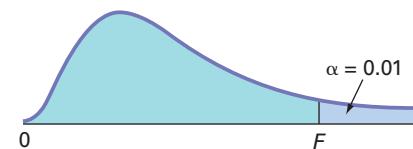


Cumulative Probabilities = 0.975																			
		Upper-Tail Areas = 0.025																	
		Numerator, df_1																	
Denominator, df_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.80	799.50	864.20	899.60	921.80	937.10	948.20	956.70	963.30	968.60	976.70	984.90	993.10	997.20	1,001.00	1,006.00	1,010.00	1,014.00	1,018.00
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.39	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
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

TABLE E.5

Critical Values of F (continued)

For a particular combination of numerator and denominator degrees of freedom, entry represents the critical values of F corresponding to the cumulative probability $(1 - \alpha)$ and a specified upper-tail area (α) .

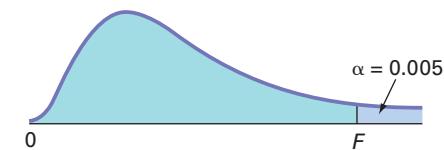


Cumulative Probabilities = 0.99																			
		Upper-Tail Areas = 0.01																	
		Numerator, df_1																	
Denominator, df_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4,052.00	4,999.50	5,403.00	5,625.00	5,764.00	5,859.00	5,928.00	5,982.00	6,022.00	6,056.00	6,106.00	6,157.00	6,209.00	6,235.00	6,261.00	6,287.00	6,313.00	6,339.00	6,366.00
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
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	2.81	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

(continued)

TABLE E.5Critical Values of F (continued)

For a particular combination of numerator and denominator degrees of freedom, entry represents the critical values of F corresponding to the cumulative probability $(1 - \alpha)$ and a specified upper-tail area (α).



Cumulative Probabilities = 0.995																			
Upper – Tail Areas = 0.005																			
		Numerator, df_1																	
Denominator, df_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	16,211.00	20,000.00	21,615.00	22,500.00	23,056.00	23,437.00	23,715.00	23,925.00	24,091.00	24,224.00	24,426.00	24,630.00	24,836.00	24,910.00	25,044.00	25,148.00	25,253.00	25,359.00	25,465.00
2	198.50	199.00	199.20	199.20	199.30	199.30	199.40	199.40	199.40	199.40	199.40	199.40	199.40	199.50	199.50	199.50	199.50	199.50	
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	12.11
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	4.61
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	4.75	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	3.41
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	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.02	3.88	3.77	3.60	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

TABLE E.6

Lower and Upper
Critical Values, T_1 , of the
Wilcoxon Rank Sum Test

n_2	α		n_1						
	One-tail	Two-tail	4	5	6	7	8	9	10
4	0.05	0.10	11,25						
	0.025	0.05	10,26						
	0.01	0.02	—,—						
	0.005	0.01	—,—						
5	0.05	0.10	12,28	19,36					
	0.025	0.05	11,29	17,38					
	0.01	0.02	10,30	16,39					
	0.005	0.01	—,—	15,40					
6	0.05	0.10	13,31	20,40	28,50				
	0.025	0.05	12,32	18,42	26,52				
	0.01	0.02	11,33	17,43	24,54				
	0.005	0.01	10,34	16,44	23,55				
7	0.05	0.10	14,34	21,44	29,55	39,66			
	0.025	0.05	13,35	20,45	27,57	36,69			
	0.01	0.02	11,37	18,47	25,59	34,71			
	0.005	0.01	10,38	16,49	24,60	32,73			
8	0.05	0.10	15,37	23,47	31,59	41,71	51,85		
	0.025	0.05	14,38	21,49	29,61	38,74	49,87		
	0.01	0.02	12,40	19,51	27,63	35,77	45,91		
	0.005	0.01	11,41	17,53	25,65	34,78	43,93		
9	0.05	0.10	16,40	24,51	33,63	43,76	54,90	66,105	
	0.025	0.05	14,42	22,53	31,65	40,79	51,93	62,109	
	0.01	0.02	13,43	20,55	28,68	37,82	47,97	59,112	
	0.005	0.01	11,45	18,57	26,70	35,84	45,99	56,115	
10	0.05	0.10	17,43	26,54	35,67	45,81	56,96	69,111	82,128
	0.025	0.05	15,45	23,57	32,70	42,84	53,99	65,115	78,132
	0.01	0.02	13,47	21,59	29,73	39,87	49,103	61,119	74,136
	0.005	0.01	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* (Pearl River, NY: Lederle Laboratories, 1964), with permission of the American Cyanamid Company.

TABLE E.7Critical Values of the Studentized Range, Q

Denominator, <i>df</i>	Upper 5% Points ($\alpha = 0.05$)																		
	Numerator, <i>df</i>																		
2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	18.00	27.00	32.80	37.10	40.40	43.10	45.40	47.40	49.10	50.60	52.00	53.20	54.30	55.40	56.30	57.20	58.00	58.80	59.60
2	6.09	8.30	9.80	10.90	11.70	12.40	13.00	13.50	14.00	14.40	14.70	15.10	15.40	15.70	15.90	16.10	16.40	16.60	16.80
3	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	9.72	9.95	10.15	10.35	10.52	10.69	10.84	10.98	11.11	11.24
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.03	8.21	8.37	8.52	8.66	8.79	8.91	9.03	9.13	9.23
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21
6	3.46	4.34	4.90	5.31	5.63	5.89	6.12	6.32	6.49	6.65	6.79	6.92	7.03	7.14	7.24	7.34	7.43	7.51	7.59
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43	6.55	6.66	6.76	6.85	6.94	7.02	7.09	7.17
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87
9	3.20	3.95	4.42	4.76	5.02	5.24	5.43	5.60	5.74	5.87	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.64
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.03	6.11	6.20	6.27	6.34	6.40	6.47
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.99	6.06	6.14	6.20	6.26	6.33
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.40	5.51	5.62	5.71	5.80	5.88	5.95	6.03	6.09	6.15	6.21
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79	5.86	5.93	6.00	6.05	6.11
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.55	5.64	5.72	5.79	5.85	5.92	5.97	6.03
15	3.01	3.67	4.08	4.37	4.60	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.58	5.65	5.72	5.79	5.85	5.90	5.96
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.52	5.59	5.66	5.72	5.79	5.84	5.90
17	2.98	3.63	4.02	4.30	4.52	4.71	4.86	4.99	5.11	5.21	5.31	5.39	5.47	5.55	5.61	5.68	5.74	5.79	5.84
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14	5.23	5.32	5.39	5.46	5.53	5.59	5.65	5.70	5.75
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43	5.49	5.55	5.61	5.66	5.71
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.25	5.32	5.38	5.44	5.50	5.54	5.59
30	2.89	3.49	3.84	4.10	4.30	4.46	4.60	4.72	4.83	4.92	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.48
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.74	4.82	4.91	4.98	5.05	5.11	5.16	5.22	5.27	5.31	5.36
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.94	5.00	5.06	5.11	5.16	5.20	5.24
120	2.80	3.36	3.69	3.92	4.10	4.24	4.36	4.48	4.56	4.64	4.72	4.78	4.84	4.90	4.95	5.00	5.05	5.09	5.13
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62	4.68	4.74	4.80	4.85	4.89	4.93	4.97	5.01

TABLE E.7Critical Values of the Studentized Range, Q (continued)

Denominator, <i>df</i>	Upper 1% Points ($\alpha = 0.01$)																		
	Numerator, <i>df</i>																		
2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	90.03	135.00	164.30	185.60	202.20	215.80	227.20	237.00	245.60	253.20	260.00	266.20	271.80	277.00	281.80	286.30	290.40	294.30	298.00
2	14.04	19.02	22.29	24.72	26.63	28.20	29.53	30.68	31.69	32.59	33.40	34.13	34.81	35.43	36.00	36.53	37.03	37.50	37.95
3	8.26	10.62	12.17	13.33	14.24	15.00	15.64	16.20	16.69	17.13	17.53	17.89	18.22	18.52	18.81	19.07	19.32	19.55	19.77
4	6.51	8.12	9.17	9.96	10.58	11.10	11.55	11.93	12.27	12.57	12.84	13.09	13.32	13.53	13.73	13.91	14.08	14.24	14.40
5	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48	10.70	10.89	11.08	11.24	11.40	11.55	11.68	11.81	11.93
6	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.49	9.65	9.81	9.95	10.08	10.21	10.32	10.43	10.54
7	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55	8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65
8	4.75	5.64	6.20	6.63	6.96	7.24	7.47	7.68	7.86	8.03	8.18	8.31	8.44	8.55	8.66	8.76	8.85	8.94	9.03
9	4.60	5.43	5.96	6.35	6.66	6.92	7.13	7.32	7.50	7.65	7.78	7.91	8.03	8.13	8.23	8.33	8.41	8.50	8.57
10	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.06	7.21	7.36	7.49	7.60	7.71	7.81	7.91	7.99	8.08	8.15	8.23
11	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13	7.25	7.36	7.47	7.56	7.65	7.73	7.81	7.88	7.95
12	4.32	5.04	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94	7.06	7.17	7.26	7.36	7.44	7.52	7.59	7.66	7.73
13	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79	6.90	7.01	7.10	7.19	7.27	7.35	7.42	7.49	7.55
14	4.21	4.90	5.32	5.63	5.88	6.09	6.26	6.41	6.54	6.66	6.77	6.87	6.96	7.05	7.13	7.20	7.27	7.33	7.40
15	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.56	6.66	6.76	6.85	6.93	7.00	7.07	7.14	7.20	7.26
16	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46	6.56	6.66	6.74	6.82	6.90	6.97	7.03	7.09	7.15
17	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38	6.48	6.57	6.66	6.73	6.81	6.87	6.94	7.00	7.05
18	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31	6.41	6.50	6.58	6.66	6.73	6.79	6.85	6.91	6.97
19	4.05	4.67	5.05	5.33	5.55	5.74	5.89	6.02	6.14	6.25	6.34	6.43	6.51	6.59	6.65	6.72	6.78	6.84	6.89
20	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19	6.29	6.37	6.45	6.52	6.59	6.65	6.71	6.77	6.82
24	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02	6.11	6.19	6.26	6.33	6.39	6.45	6.51	6.56	6.61
30	3.89	4.46	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85	5.93	6.01	6.08	6.14	6.20	6.26	6.31	6.36	6.41
40	3.83	4.37	4.70	4.93	5.11	5.27	5.39	5.50	5.60	5.69	5.76	5.84	5.90	5.96	6.02	6.07	6.12	6.17	6.21
60	3.76	4.28	4.60	4.82	4.99	5.13	5.25	5.36	5.45	5.53	5.60	5.67	5.73	5.79	5.84	5.89	5.93	5.97	6.02
120	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.38	5.44	5.51	5.56	5.61	5.66	5.71	5.75	5.79	5.83
∞	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23	5.29	5.35	5.40	5.45	5.49	5.54	5.57	5.61	5.65

Source: Extracted from H. L. Harter and D. S. Clemm, "The Probability Integrals of the Range and of the Studentized Range—Probability Integral, Percentage Points, and Moments of the Range," *Wright Air Development Technical Report 58-484*, Vol. 1, 1959.

TABLE E.8

Critical Values, d_L and d_U , of the Durbin–Watson Statistic, D (Critical Values Are One-Sided)^a

$\alpha = 0.05$										$\alpha = 0.01$										
$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$		$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$		
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U										
15	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56	2.21	.81	1.07	.70	1.25	.59	1.46	.49	1.70	.39	1.96
16	1.10	1.37	.98	1.54	.86	1.73	.74	1.93	.62	2.15	.84	1.09	.74	1.25	.63	1.44	.53	1.66	.44	1.90
17	1.13	1.38	1.02	1.54	.90	1.71	.78	1.90	.67	2.10	.87	1.10	.77	1.25	.67	1.43	.57	1.63	.48	1.85
18	1.16	1.39	1.05	1.53	.93	1.69	.82	1.87	.71	2.06	.90	1.12	.80	1.26	.71	1.42	.61	1.60	.52	1.80
19	1.18	1.40	1.08	1.53	.97	1.68	.86	1.85	.75	2.02	.93	1.13	.83	1.26	.74	1.41	.65	1.58	.56	1.77
20	1.20	1.41	1.10	1.54	1.00	1.68	.90	1.83	.79	1.99	.95	1.15	.86	1.27	.77	1.41	.68	1.57	.60	1.74
21	1.22	1.42	1.13	1.54	1.03	1.67	.93	1.81	.83	1.96	.97	1.16	.89	1.27	.80	1.41	.72	1.55	.63	1.71
22	1.24	1.43	1.15	1.54	1.05	1.66	.96	1.80	.86	1.94	1.00	1.17	.91	1.28	.83	1.40	.75	1.54	.66	1.69
23	1.26	1.44	1.17	1.54	1.08	1.66	.99	1.79	.90	1.92	1.02	1.19	.94	1.29	.86	1.40	.77	1.53	.70	1.67
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	.93	1.90	1.04	1.20	.96	1.30	.88	1.41	.80	1.53	.72	1.66
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	.95	1.89	1.05	1.21	.98	1.30	.90	1.41	.83	1.52	.75	1.65
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	.98	1.88	1.07	1.22	1.00	1.31	.93	1.41	.85	1.52	.78	1.64
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86	1.09	1.23	1.02	1.32	.95	1.41	.88	1.51	.81	1.63
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85	1.10	1.24	1.04	1.32	.97	1.41	.90	1.51	.83	1.62
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84	1.12	1.25	1.05	1.33	.99	1.42	.92	1.51	.85	1.61
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83	1.13	1.26	1.07	1.34	1.01	1.42	.94	1.51	.88	1.61
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83	1.15	1.27	1.08	1.34	1.02	1.42	.96	1.51	.90	1.60
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82	1.16	1.28	1.10	1.35	1.04	1.43	.98	1.51	.92	1.60
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	.94	1.59
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	.95	1.59
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	.97	1.59
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	.99	1.59
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

^a n = number of observations; k = number of independent variables.

Source: Computed from TSP 4.5 based on R. W. Farebrother, "A Remark on Algorithms AS106, AS153, and AS155: The Distribution of a Linear Combination of Chi-Square Random Variables," *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 1984, 29, p. 323–333.

TABLE E.9

Control Chart Factors

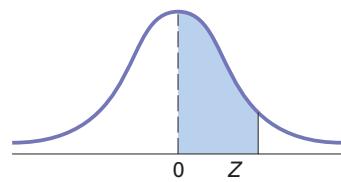
Number of Observations in Sample/Subgroup (<i>n</i>)	<i>d</i>₂	<i>d</i>₃	<i>D</i>₃	<i>D</i>₄	<i>A</i>₂
2	1.128	0.853	0	3.267	1.880
3	1.693	0.888	0	2.575	1.023
4	2.059	0.880	0	2.282	0.729
5	2.326	0.864	0	2.114	0.577
6	2.534	0.848	0	2.004	0.483
7	2.704	0.833	0.076	1.924	0.419
8	2.847	0.820	0.136	1.864	0.373
9	2.970	0.808	0.184	1.816	0.337
10	3.078	0.797	0.223	1.777	0.308
11	3.173	0.787	0.256	1.744	0.285
12	3.258	0.778	0.283	1.717	0.266
13	3.336	0.770	0.307	1.693	0.249
14	3.407	0.763	0.328	1.672	0.235
15	3.472	0.756	0.347	1.653	0.223
16	3.532	0.750	0.363	1.637	0.212
17	3.588	0.744	0.378	1.622	0.203
18	3.640	0.739	0.391	1.609	0.194
19	3.689	0.733	0.404	1.596	0.187
20	3.735	0.729	0.415	1.585	0.180
21	3.778	0.724	0.425	1.575	0.173
22	3.819	0.720	0.435	1.565	0.167
23	3.858	0.716	0.443	1.557	0.162
24	3.895	0.712	0.452	1.548	0.157
25	3.931	0.708	0.459	1.541	0.153

Source: Reprinted from *ASTM-STP 15D* by kind permission of the American Society for Testing and Materials. Copyright ASTM International, 100 Barr Harbor Drive, Conshohocken, PA 19428.

TABLE E.10

The Standardized Normal Distribution

Entry represents area under the standardized normal distribution from the mean to Z



Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.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	.49865	.49869	.49874	.49878	.49882	.49886	.49889	.49893	.49897	.49900
3.1	.49903	.49906	.49910	.49913	.49916	.49918	.49921	.49924	.49926	.49929
3.2	.49931	.49934	.49936	.49938	.49940	.49942	.49944	.49946	.49948	.49950
3.3	.49952	.49953	.49955	.49957	.49958	.49960	.49961	.49962	.49964	.49965
3.4	.49966	.49968	.49969	.49970	.49971	.49972	.49973	.49974	.49975	.49976
3.5	.49977	.49978	.49978	.49979	.49980	.49981	.49981	.49982	.49983	.49983
3.6	.49984	.49985	.49985	.49986	.49986	.49987	.49987	.49988	.49988	.49989
3.7	.49989	.49990	.49990	.49990	.49991	.49991	.49992	.49992	.49992	.49992
3.8	.49993	.49993	.49993	.49994	.49994	.49994	.49994	.49995	.49995	.49995
3.9	.49995	.49995	.49996	.49996	.49996	.49996	.49996	.49997	.49997	.49997

This appendix reviews knowledge that you will find useful if you plan to be more than a casual user of Microsoft Excel. While none of the content in this appendix needs to be mastered in order to use the instructions in the Excel Guides in this book, reviewing this appendix as necessary will help you make better sense of your Excel results. If you are using a version of Excel that is older than Excel 2010, you will need to be familiar with Section F.3 so that you can modify the names of functions used in worksheet templates and models as necessary.

Section F.4 presents an enhanced explanation of some the statistical worksheet functions that recur in two or more chapters. This section also discusses functions that either serve programming purposes or are used in novel ways to compute intermediate results. If you have a particular interest in developing application solutions, you will want to be familiar with this set of functions.

This appendix assumes that you are familiar with Excel and have mastered the basic concepts presented in Appendix B. If you are a first-time user of Excel, do not make the mistake of trying to comprehend the material in this appendix before you gain experience using Excel and familiarity with Appendix B.

F.1 Useful Keyboard Shortcuts

In Microsoft Office programs including Microsoft Excel, certain individual keys or combinations of keys held down as you press another key are shortcuts that allow you to execute common operations without having to select choices from menus or click in the Ribbon. As first explained in Table GS.2 on page 7, in this book, keystroke combinations are shown using plus signs; for example, **Ctrl+C** means “while holding down the **Ctrl** key, press the **C** key.”

Editing Shortcuts

Pressing **Backspace** erases typed characters to the left of the current position, one character at a time. Pressing **Delete** erases characters to the right of the cursor, one character at a time.

Ctrl+C copies a worksheet entry, and **Ctrl+V** pastes that entry into the place that the editing cursor or worksheet cell highlight indicates. Pressing **Ctrl+X** cuts the currently selected entry or object so that you can paste it somewhere else. **Ctrl+C** and **Ctrl+V** (or **Ctrl+X** and **Ctrl+V**) can also be used to copy (or cut) and paste certain workbook objects such as charts. (Using copy and paste to copy formulas from one worksheet cell to another is subject to the adjustment discussed in Section B-2.)

Pressing **Ctrl+Z** undoes the last operation, and **Ctrl+Y** redoing the last operation. Pressing **Enter** or **Tab** finalizes an entry typed into a worksheet cell. Pressing either key is implied by the use of the verb *enter* in the Excel Guides.

Formatting Shortcuts

Pressing **Ctrl+B** toggles on (or off) boldface text style for the currently selected object. Pressing **Ctrl+I** toggles on (or off) italic text style for the currently selected object. Pressing **Ctrl+Shift+%** formats numeric values as a percentage with no decimal places.

Utility Shortcuts

Pressing **Ctrl+F** finds a **Find what** value, and pressing **Ctrl+H** replaces a **Find what** value with the **Replace with** value. Pressing **Ctrl+A** selects the entire current worksheet (useful as part of a worksheet copy or format operation). Pressing **Esc** cancels an action or a dialog box. Pressing **F1** displays the Microsoft Excel help system.

F.2 Verifying Formulas and Worksheets

If you use formulas in your worksheets, you should review and verify formulas before you use their results. To view the formulas in a worksheet, press **Ctrl+`** (grave accent key). To restore the original view, the results of the formulas, press **Ctrl+`** a second time.

As you create and use more complicated worksheets, you might want to visually examine the relationships among a formula and the cells it uses (called the *precedents*) and the cells that use the results of the formula (the *dependents*). Select **Formulas → Trace Precedents** (or **Trace Dependents**) to examine relationships. When you are finished, clear all trace arrows by selecting **Formulas → Remove Arrows**.

After verifying formulas, you should test, using simple numbers, any worksheet that you have modified or constructed from scratch.

F.3 New Function Names

Beginning in Excel 2010, Microsoft renamed many statistical functions and reprogrammed a number of functions to improve their accuracy. Generally, with exceptions noted, this book uses the new function names in worksheet cell formulas. The new function names used in this book are listed in Table F.1, along with the place of first mention in this book and corresponding older function name.

TABLE F.1

New Function Names Used in This Book and Older ("Compatible") Names

New Name	First Mention	Older Name
BINOM.DIST	EG5.3	BINOMDIST
CHISQ.DIST.RT	EG12.1	CHIDIST
CHISQ.INV.RT	EG12.1	CHIINV
CONFIDENCE.NORM	EG8.1	CONFIDENCE
COVARIANCE.S	EG3.5	none*
EXPON.DIST	EG6.5	EXPONDIST
F.DIST.RT	EG10.4	FDIST
F.INV.RT	EG10.4	FINV
HYPGEOM.DIST	EG5.5	HYPGEOMDIST
NORM.DIST	EG6.2	NORMDIST
NORM.INV	EG6.2	NORMINV
NORM.S.DIST	EG9.1	NORMSDIST
NORM.S.INV	EG6.2	NORMSINV
POISSON.DIST	EG5.4	POISSON
STDEV.S	EG3.2	STDEV
STDEV.P	EG3.2	STDEVP
T.DIST.RT	EG9.3	TDIST
T.DIST.2T	EG9.2	TDIST
T.INV.2T	EG8.2	TINV
VAR.S	EG3.2	VAR
VAR.P	EG3.2	VARP

*COVARIANCE.S is a function that was new to Excel 2010. The COVARIANCE.P function (not used in this book) replaces the older COVAR function.

Because the new function names are not compatible with Excel versions older than Excel 2010, alternative worksheets have been included in the Excel Guide workbooks, as explained in "Alternative Worksheets," later in this section. If compatibility with older Excel versions is important to you, you should use the older function names (and the alternative worksheets).

Quartile Function

In this book, you will see the older QUARTILE function and not the newer QUARTILE.EXC function. In Microsoft's *Function Improvements in Microsoft Office Excel 2010* (available at bit.ly/RkoFIIf), QUARTILE.EXC is explained as being "consistent with industry best practices, assuming percentile is a value between 0 and 1, exclusive." Because there are several established but different ways of computing quartiles, there is no way of knowing exactly how the new function works.

Because of this lack of specifics, this book uses the older QUARTILE function, whose programming and limitations are well known, and not the new QUARTILE.EXC function or QUARTILE.INC function, which is the QUARTILES function renamed for consistency with QUARTILES.EXC. As noted in Section EG3.3, none of the three functions compute quartiles using the rules presented in Section 3.3, which are properly computed in the COMPUTE worksheet of the Quartiles workbook that uses the older QUARTILE function. If you are using Excel 2010 or a newer version of Excel, the COMPARE worksheet illustrates the results using the three forms of QUARTILES for the data found in column A of the DATA worksheet.

Alternative Worksheets

If a worksheet in an Excel Guide workbook uses one or more of the new function names, the workbook contains an alternative worksheet for use with Excel versions that are older than Excel 2010. Three exceptions to the rule are the **Simple Linear Regression 2007**, **Multiple Regression 2007**, and **Exponential Trend 2007 workbooks**. As explained in Chapters 13, 14, and 16, respectively, these workbooks serve as alternatives to the Simple Linear Regression, Multiple Regression, and Exponential Trend workbooks. Alternative worksheets and workbooks work best in Excel 2007.

The following Excel Guide workbooks contain an alternative worksheet named COMPUTE_OLDER. Numbers that appear in parentheses are the chapters in which these workbooks are first mentioned.

Parameters (3)	CIE sigma unknown (8)	Separate-Variance T (10)
Covariance (3)	CIE Proportion (8)	Paired T (10)
Hypergeometric (5)	Z Mean workbook (9)	One-Way ANOVA (11)
Normal (6)	T mean workbook (9)	Chi-Square (12)
Exponential (6)	Z Proportion (9)	
CIE sigma known (8)	Pooled-Variance T (10)	

The following Excel Guide workbooks have alternative worksheets with various names:

Descriptive (3)	CompleteStatistics_OLDER
Binomial (5)	CUMULATIVE_OLDER
Poisson (5)	CUMULATIVE_OLDER
NPP (6)	PLOT_OLDER and NORMAL_PLOT_OLDER
One-Way ANOVA (11)	TK4_OLDER
Chi-Square Worksheets (12)	Various worksheets, including ChiSquare2x3_OLDER and Marascuilo2x3_OLDER
Wilcoxon (12)	COMPUTE_ALL_OLDER
Kruskal-Wallis Worksheets (12)	KruskalWallis3_OLDER and KruskalWallis4_OLDER

As explained in Chapters 13, 14, and 16, respectively, the **Simple Linear Regression 2007**, **Multiple Regression 2007**, and **Exponential Trend 2007 workbooks** contain a number of alternative worksheets for Excel versions that are older than Excel 2010. (These alternative workbooks work best in Excel 2007.)

F.4 Understanding the Nonstatistical Functions

Although this book focuses on Excel statistical functions, selected Excel Guide and PHStat worksheets use a number of nonstatistical functions that either compute an intermediate result or perform a mathematical or programming operation. These functions are explained in the following alphabetical list:

CEILING(*cell, round-to value*) takes the numeric value in *cell* and rounds it to the next multiple of the *round-to value*. For example, if the *round-to value* is **0.5**, as it is in several column B formulas in the COMPUTE worksheet of the Quartiles workbook, then the numeric value will be rounded either to an integer or a number that contains a half such as **1.5**.

COUNT(*cell range*) counts the number of cells in a cell range that contain a numeric value. This function is often used to compute the sample size, *n*, for example, in cell B9 of the COMPUTE worksheet of the Correlation workbook. When seen in the worksheets presented in this book, the *cell range* will typically be the cell range of variable column, such as **DATA!A:A**. This will result in a proper count of the sample size of that variable if you follow the Section EG.2 rules for entering data (see pages 8 and 9).

COUNTIF(*cell range for all values, value to be matched*) counts the number of occurrences of a value in a cell range. For example, the COMPUTE worksheet of the Wilcoxon workbook uses **COUNTIF(SortedRanks!A2:A21, "Beverage")** in cell B7 to compute the sample size of the Population 1 Sample by counting the number of occurrences of the sample name Beverage in column A of the SortedRanks worksheet. In the KruskalWallis4 worksheet of the Kruskal–Wallis Worksheets workbook, the function counts the number of occurrences in column A of the SortedRanks worksheet of a supplier name that appears in a column D row.

DEVSQ(*variable cell range*) computes the sum of the squares of the differences between a variable value and the mean of that variable. For example, in Equations (3.6) on page 113 that defines the sample variance, **DEVSQ(*X variable cell range*)** computes the value of the term in the numerator of the fraction.

FLOOR(*cell, 1*) takes the numeric value in *cell* and rounds down the value to the nearest integer.

IF(*logical comparison, what to display if comparison holds, what to display if comparison is false*) uses the *logical comparison* to make a choice between two alternatives. In the worksheets shown in this book, the IF function typically chooses from two text values, such as **Reject the null hypothesis** and **Do not reject the null hypothesis**, to display, but in Chapter 16, the function is used to create dummy variables for quarterly or monthly data.

MMULT(*cell range 1, cell range 2*) treats both *cell range 1* and *cell range 2* as matrices and computes the matrix product of the two matrices. When each of the two cell ranges is either a single row or a single column, MMULT can be used as part of a regular formula. If the cell ranges each represent rows and columns, then MMULT must be used as part of an array formula (see Appendix Section B.3). One exception to these rules occurs in cell B21 of the CIEandPI worksheet of the Multiple Regression worksheet, in which **MMULT(TRANSPOSE(B5:B7), COMPUTE!B17:B19)** has been entered as part of an array formula because of how Excel treats the results of the TRANSPOSE function.

ROUND(*cell, 0*) takes the numeric value in *cell* and rounds to the nearest whole number.

SMALL(*cell range, k*) selects the *kth* smallest value in *cell range*.

SQRT(*value*) computes the square root of *value*, where *value* is either a cell reference or an arithmetic expression.

SUMIF(*cell range for all values, value to be matched, cell range in which to select cells for summing*) sums only those rows in *cell range in which to select cells for summing* in which the value in *cell range for all values* matches the *value to be matched*. SUMIF provides a convenient way to compute the sum of ranks for a sample in a worksheet that contains stacked data. For example, the COMPUTE worksheet of the Wilcoxon workbook uses **SUMIF(SortedRanks!A2:A21, "Beverage", SortedRanks!C2:C21)** in cell B8 to compute the sum of ranks for the Beverage (End-cap) sample by summing only the rows in the SortedRanks worksheet column C whose column A value is Beverage. In the KruskalWallis4

worksheet of the Kruskal–Wallis Worksheets workbook, SUMIF sums only the rows in the SortedRanks worksheet column C whose column A value matches the value that appears in a column D row.

SUMPRODUCT(*cell range 1, cell range 2*) multiplies each cell in *cell range 1* by the corresponding cell in *cell range 2* and then sums those products. If *cell range 1* contains a column of differences between an *X* value and the mean of the variable *X*, and *cell range 2* contains a column of differences between a *Y* value and the mean of the variable *Y*, then this function would compute the value of the numerator in Equation (3.16) that defines the sample covariance. In Section EG16.6, **SUMPRODUCT(ABS(*cell range of residual values*))** uses the function in a novel way with only one cell range to efficiently compute the sum of the absolute values of the values found in the *cell range of residual values*.

TRANSPOSE(*horizontal or vertical cell range*) takes the *cell range*, which must be either a horizontal cell range (cells all in the same row) or a vertical cell range (cells all in the same column) and transposes, or rearranges, the cell in the other orientation such that a horizontal cell range becomes a vertical cell range and vice versa. When used inside another function, Excel considers the results of this function to be an *array*, not a cell range.

VLOOKUP(*lookup value cell, table of lookup values, table column to use*) function displays a value that has been looked up in a *table of lookup values*, a rectangular cell range. In the ADVANCED worksheet of the Recoded workbook, the function uses the values in the second column of *table of lookup values* (an example of which is shown below) to look up the Honors values based on the GPA of a student (the *lookup value cell*). Numbers in the first column of *table of lookup values* are implied ranges such that No Honors is the value displayed if the GPA is at least 0, but less than 3; Honor Roll is the value displayed if the GPA is at least 3, but less than 3.3; and so on:

0	No Honors
3	Honor Roll
3.3	Dean's List
3.7	President's List

APPENDIX G Software FAQs

G.1 PHStat FAQs

What is PHStat?

PHStat is the macro-enabled workbook that you use with Excel to help build solutions to statistical problems. With PHStat, you fill in simple-to-use dialog boxes and watch as PHStat creates a worksheet solution for you. PHStat allows you to use the Microsoft Excel statistical functions without having to first learn advanced Excel techniques or worrying about building worksheets from scratch. As a student studying statistics, you can focus mainly on learning statistics and not worry about having to fully master Excel as well.

PHStat executes for you the low-level menu selection and worksheet entry tasks that are associated with implementing statistical analysis in Microsoft Excel. PHStat creates worksheets and chart sheets that are identical to the ones featured in this book. From these sheets, you can learn real Excel techniques at your leisure and give yourself the ability to use Excel effectively outside your introductory statistics course. (Other add-ins that appear similar to PHStat report results as a series of text labels, hiding the details of using Microsoft Excel and leaving you with no basis for learning to use Excel effectively.)

Which versions of Excel are compatible with PHStat?

PHStat works best with Microsoft Windows Excel 2010 and Excel 2013 and with OS X Excel 2011. PHStat is also compatible with Excel 2007 (WIN), although the accuracy of some Excel statistical functions PHStat uses varies from Excel 2010 and can lead to (minor) changes in the results reported.

PHStat is partially compatible with Excel 2003 (WIN). When you open PHStat in Excel 2003, you will see a file conversion dialog box as Excel translates the .xlam file into a format that can be used in Excel 2003. After this file conversion completes, you will be able to see the PHStat menu and use many of the PHStat procedures. As documented in the PHStat help system some advanced procedures construct worksheets that use Excel functions that were added after Excel 2003 was published. In those cases, the worksheets will contain cells that display the #NAME? error message instead of results.

PHStat is not compatible with Excel 2008 (OS X), which did include the capability of running add-in workbooks.

How do I get PHStat ready for use?

Section D.2 explains how to get PHStat ready for use. You should also review the PHStat readme file (available

for download as discussed in Appendix C) for any late-breaking news or changes that might affect this process.

When I open PHStat, I get a Microsoft Excel error message that mentions a “compile error” or “hidden workbook.” What is wrong?

Most likely, you have not applied the Microsoft-supplied updates to your copy of Microsoft Excel (see Section D.1). If you are certain that your copy of Microsoft Excel is fully up-to-date, verify that your copy is properly licensed and undamaged. (If necessary, you can rerun the Microsoft Office setup program to repair the installation of Excel.)

When I use a particular PHStat procedure, I get an error message that includes the words “unexpected error.” What should I do?

“Unexpected error” messages are typically caused by improperly prepared data. Review your data to ensure that you have organized your data according to the conventions PHStat expects, as explained in the Section EG.5 on page 12 and the PHStat help system, and “clean” your data, as discussed in Section 1.3, if necessary.

Where can I get further news and information about PHStat? Where can I get further assistance about using PHStat?

Several websites can provide you with news and information or provide you with assistance that supplements the readme file and help system included with PHStat.

www.pearsonhighered.com/phstat is Pearson Education’s official web page for PHStat. From this page, you can download PHStat (requires an access code as explained in Section C.4) or contact Pearson 24/7 Technical Support directly about any technical issue that you cannot resolve.

phstat.davidlevinestatistics.com is a website maintained by the authors of this book that contains general news and information about PHStat.

phstatcommunity.org is a new website organized by PHStat users and endorsed by the developers of PHStat. You can click News on the home page to display the latest news and developments about PHStat. Other content on the website explains some of the “behind-the-scenes” technical workings of PHStat.

How can I make sure that my version of PHStat is up-to-date? How can I get updates to PHStat when they become available?

PHStat is subject to continuous improvement. When enhancements are made, a new PHStat zip archive is posted on the PHStat home page (see Section C.4) and, if you hold a valid

access code, you can download that archive and overwrite your older version. To discover the version number of your copy of PHStat, select **About PHStat** from the PHStat menu. (The version number for the PHStat version supplied for use with this book will always be a number that begins with 4.)

G.2 Microsoft Excel FAQs

Do all Microsoft Excel versions contain the same features and functionality? Which Microsoft Excel version should I use?

Unfortunately, features and functionality vary across versions still in use (including versions no longer supported by Microsoft). This book works best with Microsoft Windows versions Excel 2010 and Excel 2013 and OS X version Excel 2011. However, even among these current versions there are variations in features. For example, the slicer functionality discussed in Section 17.1 is found only in Excel 2010 and Excel 2013 and is missing in OS X Excel 2011 as well as in older Microsoft Windows versions. PivotTables have subtle differences across versions, none of which affect the instructions and examples in this book, and PivotCharts, not discussed in this book, are not included in Excel 2011 (see related PivotChart FAQ).

This book identifies differences among versions when they are significant. In particular, this book supplies, when necessary, special instructions and alternative worksheets (discussed in Appendix Section F.3) designed for versions that are both older than Excel 2010 and currently supported by Microsoft. If you plan to use Microsoft Windows Excel 2007, an upgrade will give you access to the newest features and provide a version with significantly increased statistical accuracy.

If you use OS X Excel 2008, you *must* upgrade to use PHStat or any of the other add-in workbooks mentioned in this book. Even if you plan to avoid using any add-ins, you should consider upgrading to OS X Excel 2011 for the same reasons that Excel 2003 and Excel 2007 face.

What does “Compatibility Mode” in the title bar mean?

Excel displays “Compatibility Mode” when you open and use a workbook that has been previously stored using the older .xls Excel workbook file format. Compatibility Mode does not affect Excel functionality but will cause Excel to review your workbook for exclusive-to-xlsx formatting properties and Excel will question you with a dialog box should you go to save the workbook in this format.

To convert a .xls workbook to the .xlsx format, select **File → Save As** and select **Excel Workbook (*.xlsx)** from the **Save as type** (WIN) or the **Format** (OS X) drop-down list in Excel 2010, 2011, or 2013. To do so in Excel 2007, click the **Office Button**, move the mouse pointer over **Save As**, and, in the Save As gallery, click **Excel Workbook** to save the workbook in the .xlsx file format.

One quirk in Microsoft Excel is that when you convert a workbook by using **Save As**, the newly converted .xlsx workbook stays temporarily in Compatibility Mode. To avoid possible complications and errors, close the newly converted workbook and then reopen it.

Using Compatibility Mode can cause minor differences in the objects such as charts and PivotTables that Excel creates and can cause problems when you seek to transfer data from other workbooks. Unless you need to open a workbook in a version of Excel that is older than Excel 2007, you should avoid using Compatibility Mode.

What Excel security settings will allow the PHStat or a Visual Explorations add-in workbook to function properly when using a Microsoft Windows version of Microsoft Excel?

The security settings are explained in the Appendix Section D.3 instructions. (These settings do not apply to OS X Excel.)

What is a PivotChart? Why doesn't this book discuss PivotCharts?

PivotCharts are charts that Microsoft Excel creates automatically from a PivotTable. This type of chart is not discussed in this book because Excel will typically create a “wrong” chart that takes more effort to fix than the effort needed to create a proper chart and because PivotChart functionality varies very significantly among the current Excel versions—and is missing from OS X Excel 2011.

The special instructions for selecting a PivotTable cell or cell range that appear in selected Section EG2.3 *In-Depth Excel* instructions help you avoid creating an unwanted PivotChart. (PHStat never creates a PivotChart.)

What is Microsoft SkyDrive?

Microsoft SkyDrive is an Internet-based service that offers you online storage that enables you to access and share your files anytime and anywhere there is an Internet connection available. In Excel 2013, you will see **SkyDrive** listed as a choice along with **Computer** in the Open, Save, and Save As panels. In Excel 2011, you use the Document Connection to access SkyDrive files and select **File → Share → Open SkyDrive** to save to a SkyDrive folder.

You must sign in to the SkyDrive service using a “Microsoft account,” formerly known as a “Windows Live ID.” If you use the Microsoft Office Web Excel app, or certain other special versions of Excel, you *may* need to sign into the SkyDrive service to use Excel itself.

What is Office 365?

Office 365 is a subscription-based service that supplies you with the latest version of Microsoft Office programs for your system. Office 365 requires you to be signed in using a Microsoft account in the same way as you would sign in to use SkyDrive (see previous answer). Using Office 365 gives you access to the latest version of Microsoft Excel, which, at the time of publication of this book, is Excel 2013 for Microsoft

Windows systems and Excel 2011 for OS X systems. If you use Office 365, use either the Excel 2013 or Excel 2011 instructions, as appropriate.

G.3 FAQs for New Users of Microsoft Excel 2013

When I open Excel 2013, I see a screen that shows panels that represent different workbooks and not the Ribbon interface. What do I do?

Press **Esc**. That screen, called the **Start screen**, will disappear and a screen that contains an Excel window similar to the ones in Excel 2010 and Excel 2011 will appear. For a more permanent solution, select **File → Options** and in the General panel of the Excel Options dialog box that appears clear **Show the Start screen when this application starts** and then click **OK**.

Are there any significant differences between Excel 2013 and its immediate predecessor, Excel 2010?

There are no significant differences, but several File tab commands present restyled panes (with the same or similar information), and opening and saving files differs slightly, as described in the Excel Guide for the Let's Get Started chapter.

The Excel 2013 Ribbon, featured in a number of Appendix B illustrations, looks slightly different than the Excel 2010 Ribbon. However, these differences are so slight that the Excel 2013 Ribbon illustrations in Appendix B will be recognizable to you if you choose to use Excel 2010. The Excel 2013 Ribbon also contains a number of new icons and groups in some of its tabs, but those additions do not affect any of the Ribbon selection sequences presented in the Excel Guides.

In the Insert tab, what are Recommended PivotTables and Recommended Charts? Should I use these features?

Recommended PivotTables and Recommended Charts display one or more “recommended” PivotTables or charts as shortcuts. Unfortunately, the recommended PivotTables can include statistical errors such as treating the categories of a categorical variable as zero values of a numerical variable and the recommended charts often do not conform to best practices (see Appendix Section B.6).

As programmed in Excel 2013, you should ignore and not use these features as they will likely cause you to spend more time correcting errors and formatting mistakes than the little time that you might otherwise save.

G.4 Minitab FAQs

Can I use Minitab Release 14 or 15 with this book?

Yes, you can use the Minitab Guide instructions, written for Minitab 16, with Release 14 or 15. For certain methods, there may be minor differences in labeling of dialog box elements. Any difference that is not minor is noted in the instructions.

Can I save my Minitab worksheets or projects for use with Release 14 or 15?

Yes. Select either **Minitab14** or **Minitab 15** (for a worksheet) or **Minitab 14 Project (*.MPJ)** or **Minitab 15 Project (*.MPJ)** (for a project) from the **Save as type** drop-down list in the save as dialog box. See Section MG1.3 on page 53 for more information about using the Save Worksheet As and Save Project As dialog boxes.

Self-Test Solutions and Answers to Selected Even-Numbered Problems

The following sections present worked-out solutions to Self-Test Problems and brief answers to most of the even-numbered problems in the text. For more detailed solutions, including explanations, interpretations, and Excel and Minitab results, see the *Student Solutions Manual*.

CHAPTER 1

1.2 Small, medium, and large sizes imply order but do not specify how much more soft drink is added at increasing levels.

1.4 (a) The number of cellphones is a numerical variable that is discrete because the outcome is a count. It is ratio scaled because it has a true zero point. **(b)** Monthly data usage is a numerical variable that is continuous because any value within a range of values can occur. It is ratio scaled because it has a true zero point. **(c)** Number of text messages exchanged per month is a numerical variable that is discrete because the outcome is a count. It is ratio scaled because it has a true zero point. **(d)** Voice usage per month is a numerical variable that is continuous because any value within a range of values can occur. It is ratio scaled because it has a true zero point. **(e)** Whether a cellphone is used for email is a categorical variable because the answer can be only yes or no. This also makes it a nominal-scaled variable.

1.6 (a) Categorical, nominal scale. **(b)** Numerical, continuous, ratio scale. **(c)** Categorical, nominal scale. **(d)** Numerical, discrete, ratio scale. **(e)** Categorical, nominal scale.

1.8 (a) Numerical, continuous, ratio scale. **(b)** Numerical, discrete, ratio scale. **(c)** Numerical, continuous, ratio scale. **(d)** Categorical, nominal scale.

1.10 The underlying variable, ability of the students, may be continuous, but the measuring device, the test, does not have enough precision to distinguish between the two students.

1.18 Sample without replacement: Read from left to right in three-digit sequences and continue unfinished sequences from the end of the row to the beginning of the next row:

Row 05: 338 505 855 551 438 855 077 186 579 488 767 833 170

Rows 05–06: 897

Row 06: 340 033 648 847 204 334 639 193 639 411 095 924

Rows 06–07: 707

Row 07: 054 329 776 100 871 007 255 980 646 886 823 920 461

Row 08: 893 829 380 900 796 959 453 410 181 277 660 908 887

Rows 08–09: 237

Row 09: 818 721 426 714 050 785 223 801 670 353 362 449

Rows 09–10: 406

Note: All sequences above 902 and duplicates are discarded.

1.20 A simple random sample would be less practical for personal interviews because of travel costs (unless interviewees are paid to go to a central interviewing location).

1.22 Here all members of the population are equally likely to be selected, and the sample selection mechanism is based on chance. But selection of two elements is not independent; for example, if *A* is in the sample, we know that *B* is also and that *C* and *D* are not.

1.24 (a)

Row 16: 2323 6737 5131 8888 1718 0654 6832 4647 6510 4877

Row 17: 4579 4269 2615 1308 2455 7830 5550 5852 5514 7182

Row 18: 0989 3205 0514 2256 8514 4642 7567 8896 2977 8822

Row 19: 5438 2745 9891 4991 4523 6847 9276 8646 1628 3554

Row 20: 9475 0899 2337 0892 0048 8033 6945 9826 9403 6858

Row 21: 7029 7341 3553 1403 3340 4205 0823 4144 1048 2949

Row 22: 8515 7479 5432 9792 6575 5760 0408 8112 2507 3742

Row 23: 1110 0023 4012 8607 4697 9664 4894 3928 7072 5815

Row 24: 3687 1507 7530 5925 7143 1738 1688 5625 8533 5041

Row 25: 2391 3483 5763 3081 6090 5169 0546

Note: All sequences above 5,000 are discarded. There were no repeating sequences.

(b)

089	189	289	389	489	589	689	789	889	989
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

1089	1189	1289	1389	1489	1589	1689	1789	1889	1989
------	------	------	------	------	------	------	------	------	------

2089	2189	2289	2389	2489	2589	2689	2789	2889	2989
------	------	------	------	------	------	------	------	------	------

3089	3189	3289	3389	3489	3589	3689	3789	3889	3989
------	------	------	------	------	------	------	------	------	------

4089	4189	4289	4389	4489	4589	4689	4789	4889	4989
------	------	------	------	------	------	------	------	------	------

(c) With the single exception of invoice 0989, the invoices selected in the simple random sample are not the same as those selected in the systematic sample. It would be highly unlikely that a simple random sample would select the same units as a systematic sample.

1.26 Before accepting the results of a survey of college students, you might want to know, for example: Who funded the survey? Why was it conducted? What was the population from which the sample was selected? What sampling design was used? What mode of response was used: a personal interview, a telephone interview, or a mail survey? Were interviewers trained? Were survey questions field-tested? What questions were asked? Were the questions clear, accurate, unbiased, and valid? What operational definition of “vast majority” was used? What was the response rate? What was the sample size?

1.28 The results are based on an online survey. If the frame is supposed to be smartphone and tablet users, how is the population defined? This is a self-selecting sample of people who responded online, so there is an undefined nonresponse error. Sampling error cannot be determined since this is not a random sample.

1.30 Before accepting the results of the survey, you might want to know, for example: Who funded the study? Why was it conducted? What was the population from which the sample was selected? What sampling design was used? What mode of response was used: a personal interview, a telephone interview, or a mail survey? Were interviewers trained? Were survey questions field-tested? What other questions were asked? Were the questions clear, accurate, unbiased, and valid? What was the response rate? What was the margin of error? What was the sample size? What frame was used?

1.44 (a) All benefitted employees at the university. **(b)** The 3,095 employees who responded to the survey. **(c)** Gender and marital status are categorical. Age (years), education level (years completed), and household income (\$) are numerical.

CHAPTER 2

2.2 (a) Table of frequencies for all student responses:

STUDENT MAJOR CATEGORIES

GENDER	A	C	M	Totals
Male	14	9	2	25
Female	6	6	3	15
Totals	20	15	5	40

(b) Table based on total percentages:

STUDENT MAJOR CATEGORIES

GENDER	A	C	M	Totals
Male	35.0%	22.5%	5.0%	62.5%
Female	15.0	15.0	7.5	37.5
Totals	50.0	37.5	12.5	100.0

Table based on row percentages:

STUDENT MAJOR CATEGORIES

GENDER	A	C	M	Totals
Male	56.0%	36.0%	8.0%	100.0%
Female	40.0	40.0	20.0	100.0
Totals	50.0	37.5	12.5	100.0

Table based on column percentages:

STUDENT MAJOR CATEGORIES

GENDER	A	C	M	Totals
Male	70.0%	60.0%	40.0%	62.5%
Female	30.0	40.0	60.0	37.5
Totals	100.0	100.0	100.0	100.0

2.4 (a) The percentage of complaints for each automaker:

Automaker	Frequency	Percentage	Cumulative Pct.
General Motors	551	18.91%	18.91%
Other	516	17.71%	36.62%
Nissan Motors Corporation	467	16.03%	52.64%
Ford Motor Company	440	15.10%	67.74%
Chrysler LLC	439	15.07%	82.81%
Toyota Motor Sales	332	11.39%	94.20%
American Honda	169	5.80%	100.00%

(b) General Motors has the most complaints, followed by Other, Nissan Motors Corporation, Ford Motor Company, Chryler LLC, Toyota Motor Sales and American Honda.

(c) The percentage of complaints for each category:

Category	Frequency	Percentage	Cumulative Pct.
Powertrain	1148	42.82%	42.82%
Steering	397	14.81%	57.63%
Interior Electronics/Hardware	279	10.41%	68.03%
Fuel/Emission/Exhaust System	240	8.95%	76.99%
Airbags and Seatbelts	201	7.50%	84.48%
Body and Glass	182	6.79%	91.27%
Brakes	163	6.08%	97.35%
Tires and Wheels	71	2.65%	100.00%

(d) Powertrain has the most complaints, followed by steering, interior electronics/hardware, fuel/emission/exhaust system, airbags and seatbelts, body and glass, brakes, and, finally, tires and wheels.

2.6 (a) The percentages are 4.00, 10.58, 25.91, and 59.51. **(b)** More than half the oil produced is from non-OPEC countries. More than 25% is produced by OPEC countries other than Iran and Saudi Arabia.

2.8 (a) Table of row percentages:

ENJOY SHOPPING FOR CLOTHING	GENDER		
	Male	Female	Total
Yes	46%	54%	100%
No	53	47	100
Total	50	50	100

Table of column percentages:

ENJOY SHOPPING FOR CLOTHING	GENDER		
	Male	Female	Total
Yes	44%	51%	47%
No	56	49	53
Total	100	100	100

Table of total percentages:

ENJOY SHOPPING FOR CLOTHING	GENDER		
	Male	Female	Total
Yes	22%	25%	47%
No	28	25	53
Total	50	50	100

(b) A higher percentage of females enjoy shopping for clothing.

2.10 Social recommendations had very little impact on correct recall. Those who arrived at the link from a recommendation had a correct recall of 73.07% as compared to those who arrived at the link from browsing who had a correct recall of 67.96%.

2.12 73 78 78 78 85 88 91.

2.14 (a) 0 but less than 5 million, 5 million but less than 10 million, 10 million but less than 15 million, 15 million but less than 20 million, 20 million but less than 25 million, 25 million but less than 30 million.

(b) 5 million. **(c)** 2.5 million, 7.5 million, 12.5 million, 17.5 million, 22.5 million, and 27.5 million.

2.16 (a)

Electricity Costs	Frequency	Percentage
\$80 up to \$99	4	8%
\$100 up to \$119	7	14
\$120 up to \$139	9	18
\$140 up to \$159	13	26
\$160 up to \$179	9	18
\$180 up to \$199	5	10
\$200 up to \$219	3	6

(b)

Electricity Costs	Frequency	Percentage	Cumulative %
\$ 99	4	8.00%	8.00%
\$119	7	14.00	22.00
\$139	9	18.00	40.00
\$159	13	26.00	66.00
\$179	9	18.00	84.00
\$199	5	10.00	94.00
\$219	3	6.00	100.00

(c) The majority of utility charges are clustered between \$120 and \$180.

2.18 (a), (b)

Credit Score	Frequency	Percentage	Cumulative %
695 but less than 705	3	2.10%	2.10%
705 but less than 715	12	8.39%	10.49%
715 but less than 725	12	8.39%	18.88%
715 but less than 735	19	13.29%	32.17%
735 but less than 745	18	12.59%	44.76%
745 but less than 755	24	16.78%	61.54%
755 but less than 765	22	15.38%	76.92%
765 but less than 775	20	13.99%	90.91%
775 but less than 785	10	6.99%	97.90%
795 but less than 795	3	2.10%	100.00%

(c) The average credit scores are concentrated around 750.

2.20 (a)

Width	Frequency	Percentage
8.310–8.329	3	6.12%
8.330–8.349	2	4.08
8.350–8.369	1	2.04
8.370–8.389	4	8.16
8.390–8.409	5	10.20
8.410–8.429	16	32.65
8.430–8.449	5	10.20
8.450–8.469	5	10.20
8.470–8.489	6	12.24
8.490–8.509	2	4.08

(b)

Width	Percentage Less Than
8.310	0
8.330	6.12
8.350	10.20
8.370	12.24
8.390	20.40
8.410	30.60
8.430	63.25
8.450	73.45
8.470	83.65
8.490	95.89
8.51	100.00

(c) All the troughs will meet the company's requirements of between 8.31 and 8.61 inches wide.

2.22 (a)

Bulb Life (hours)	Percentage, Mfgr A	Percentage, Mfgr B
6,500–7,499	7.5%	0.0%
7,500–8,499	12.5	5.0
8,500–9,499	50.0	20.0
9,500–10,499	22.5	40.0
10,500–11,499	7.5	22.5
11,500–12,499	0.0	12.5

(b)

% Less Than	Percentage Less Than, Mfgr A	Percentage Less Than, Mfgr B
7,500	7.5%	0.0%
8,500	20.0	5.0
9,500	70.0	25.0
10,500	92.5	65.0
11,500	100.0	87.5
12,500	100.0	100.0

(c) Manufacturer B produces bulbs with longer lives than Manufacturer A. The cumulative percentage for Manufacturer B shows that 65% of its bulbs lasted less than 10,500 hours, contrasted with 92.5% of Manufacturer A's bulbs. None of Manufacturer A's bulbs lasted at least 11,500 hours, but 12.5% of Manufacturer B's bulbs lasted at least 11,500 hours. At the same time, 7.5% of Manufacturer A's bulbs lasted less than 7,500 hours, whereas none of Manufacturer B's bulbs lasted less than 7,500 hours.

2.24 (b) The Pareto chart is best for portraying these data because it not only sorts the frequencies in descending order but also provides the cumulative line on the same chart. (c) You can conclude that primary branding accounts for the largest percentage, 45%. When a mix of branding and direct response is added to primarily branding, this accounts for 84%.

2.26 (b) 86%. (d) The Pareto chart allows you to see which sources account for most of the electricity.

2.28 (b) Since energy use is spread over many types of appliances, a bar chart may be best in showing which types of appliances used the most energy. (c) Heating, water heating, and cooling accounted for 72% of the residential energy use in the United States.

2.30 (b) A higher percentage of females enjoy shopping for clothing.

2.32 (b) Social recommendations had very little impact on correct recall.

2.34 50 74 74 76 81 89 92.

2.36 (a)

Stem Unit	10	Stem Unit	10
12		23	0 3
13		24	2
14	6	25	7
15	1 2 3	26	
16	0 6	27	
17	2 2 4 6	28	
18	4 4 5 7	29	
19	6 8	30	0
20		31	
21	3 7	32	4
22	3 4 4 5 5	33	7

(b) The results are concentrated between \$172 and \$225.

2.38 (c) The majority of utility charges are clustered between \$120 and \$180.

2.40 Property taxes seem concentrated between \$1,000 and \$1,500 and also between \$500 and \$1,000 per capita. There were more states with property taxes per capita below \$1,500 than above \$1,500.

2.42 The average credit scores are concentrated around 750.

2.44 (c) All the troughs will meet the company's requirements of between 8.31 and 8.61 inches wide.

2.46 (c) Manufacturer B produces bulbs with longer lives than Manufacturer A.

2.48 (b) Yes, there is a strong positive relationship between X and Y . As X increases, so does Y .

2.50 (c) There appears to be a linear relationship between the first weekend gross and either the U.S. gross or the worldwide gross of Harry Potter movies. However, this relationship is greatly affected by the results of the last movie, *Deathly Hallows, Part II*.

2.52 (a), (c) There appears to be a positive relationship between the coaches' total pay and revenue. Yes, this is borne out by the data.

2.54 (b) There is a great deal of variation in the returns from decade to decade. Most of the returns are between 5% and 15%. The 1950s, 1980s, and 1990s had exceptionally high returns, and only the 1930s and 2000s had negative returns.

2.56 (b) There was a slight decline in movie attendance between 2001 and 2012. During that time, movie attendance increased from 2002 to 2004 but then decreased to a level below that in 2001.

2.58 (a) Pivot table of tallies in terms of counts:

	Five	Four	One	Three	Two	Grand Total
Growth	18	76	16	74	43	227
Large	9	31	5	37	21	103
Mid-Cap	7	28	4	20	13	72
Small	2	17	7	17	9	52
Value	5	22	7	36	19	89
Large	2	13	5	21	9	50
Mid-Cap	1	4		9	5	19
Small	2	5	2	6	5	20
Grand Total	23	98	23	110	62	316

Pivot table in terms of % of total

	Five	Four	One	Three	Two	Grand Total
Growth	5.70%	24.05%	5.06%	23.42%	13.61%	71.84%
Large	2.85%	9.81%	1.58%	11.71%	6.65%	32.59%
Mid-Cap	2.22%	8.86%	1.27%	6.33%	4.11%	22.78%
Small	0.63%	5.38%	2.22%	5.38%	2.85%	16.46%
Value	1.58%	6.96%	2.22%	11.39%	6.01%	28.16%
Large	0.63%	4.11%	1.58%	6.65%	2.85%	15.82%
Mid-Cap	0.32%	1.27%	0.00%	2.85%	1.58%	6.01%
Small	0.63%	1.58%	0.63%	1.90%	1.58%	6.33%
Grand Total	7.28%	31.01%	7.28%	34.81%	19.62%	100.00%

(b) Patterns of star rating conditioned on market cap:

For the growth funds as a group, most are rated as four-star, followed by three-star, two-star, five-star, and one-star. The pattern of star rating is the same across the different market caps within the growth funds with most of the funds receiving a four-star rating, followed by three-star, two-star, five-star, and one-star with the exception of small-cap funds with most of the funds receiving a four-star or three-star rating, followed by two-star, one-star, and five-star.

For the value funds as a group, most are rated as three-star, followed by four-star, two-star, one-star, and five-star. Within the value funds, the large-cap funds follow the same pattern as the value funds as a group. Most of the mid-cap funds are rated as three-star, followed by two-star, four-star, five-star, and one-star while most of the small-cap funds are rated as three-star, followed by either two-star or four-star, and either one-star or five star.

Patterns of market cap conditioned on star rating:

Most of the growth funds are large-cap, followed by mid-cap and small-cap. The pattern is similar among the five-star, four-star, three-star, and two-star growth funds, but among the one-star growth funds, most are small-cap, followed by large-cap and mid-cap.

The largest share of the value funds is large-cap, followed by small-cap and mid-cap. The pattern is similar among the four-star and one-star value funds. Among the three-star value funds, most are large-cap, followed by mid-cap and then small-cap while most are large-cap, followed by equal portions of mid-cap and small-cap among the two-star value funds and most are either large-cap or small-cap followed by mid-cap among the five-star value funds.

2.60 (a) Pivot table of tallies in terms of counts:

	Five	Four	One	Three	Two	Grand Total
Growth	18	76	16	74	43	227
Average	3	15	6	28	22	74
High		1	5	1	3	10
Low	15	60	5	45	18	143
Value	5	22	7	36	19	89
Average	1		3	7	6	17
High			2		1	3
Low	4	22	2	29	12	69
Grand Total	23	98	23	110	62	316

Pivot table of tallies in terms of percentage of grand total:

	Five	Four	One	Three	Two	Grand Total
Growth	5.70%	24.05%	5.06%	23.42%	13.61%	71.84%
Average	0.95%	4.75%	1.90%	8.86%	6.96%	23.42%
High	0.00%	0.32%	1.58%	0.32%	0.95%	3.16%
Low	4.75%	18.99%	1.58%	14.24%	5.70%	45.25%
Value	1.58%	6.96%	2.22%	11.39%	6.01%	28.16%
Average	0.32%	0.00%	0.95%	2.22%	1.90%	5.38%
High	0.00%	0.00%	0.63%	0.00%	0.32%	0.95%
Low	1.27%	6.96%	0.63%	9.18%	3.80%	21.84%
Grand Total	7.28%	31.01%	7.28%	34.81%	19.62%	100.00%

(b) Patterns of star rating conditioned on risk:

For the growth funds as a group, most are rated as four-star, followed by three-star, two-star, five-star, and one-star. The pattern of star rating is the same among the low-risk growth funds. The pattern is different among the high-risk and average-risk growth funds. Among the

high-risk growth funds, most are rated as one-star, followed by two-star, equal portions of three-star and four-star, with no five-star. Among the average-risk growth funds, most are rated as three-star, followed by two-star, four-star, one-star, and five-star.

For the value funds as a group, most are rated as three-star, followed by four-star, two-star, one-star and five-star. Among the average-risk value funds, most are three-star, followed by two-star, five-star, and one-star with no four-star. Among the high-risk value funds, most are one-star, followed by two-star with no three-star, four-star, or five-star. Among the low-risk value funds, most are three-star, followed by four-star, two-star, five-star, and one-star.

Patterns of risk conditioned on star rating:

Most of the growth funds are rated as low-risk, followed by average-risk and then high-risk. The pattern is the same among the three-star, four-star, and five-star growth funds. Among the one-star growth funds, most are average-risk, followed by equal portions of high-risk and low-risk. Among the two-star growth funds, most are average-risk, followed by low-risk and high-risk.

Most of the value funds are rated as low-risk, followed by average-risk and then high-risk. The pattern is the same among the two-star, three-star, and five-star value funds. Among the one-star value funds, most are average-risk, followed by equal portions of high-risk and low-risk. Among the four-star value funds, all are low-risk with no average-risk or high-risk.

2.80 (c) The publisher gets the largest portion (64.8%) of the revenue. About half (32.3%) of the revenue received by the publisher covers manufacturing costs. The publisher's marketing and promotion account for the next largest share of the revenue, at 15.4%. Author, bookstore employee salaries and benefits, and publisher administrative costs and taxes each account for around 10% of the revenue, whereas the publisher after-tax profit, bookstore operations, bookstore pretax profit, and freight constitute the "trivial few" allocations of the revenue. Yes, the bookstore gets twice the revenue of the authors.

2.82 (b) The pie chart may be best since with only three categories, it enables you to see the portion of the whole in each category. **(d)** The pie chart may be best since, with only four categories it enables you to see the portion of the whole in each category. **(e)** The online content is not copy-edited or fact-checked as carefully as print content. Only 41% of the online content is copy-edited as carefully as print content, and only 57% of the online content is fact-checked as carefully as the print content.

2.84 (a)

DESSERT ORDERED	GENDER		
	Male	Female	Total
Yes	34%	66%	100%
No	52	48	100
Total	48	52	100

DESSERT ORDERED	GENDER		
	Male	Female	Total
Yes	17%	29%	23%
No	83	71	77
Total	100	100	100

DESSERT ORDERED	GENDER		
	Male	Female	Total
Yes	8%	15%	23%
No	40	37	77
Total	48	52	100

DESSERT ORDERED	BEEF ENTRÉE		
	Yes	No	Total
Yes	52%	48%	100%
No	25	75	100
Total	31	69	100

DESSERT ORDERED	BEEF ENTRÉE		
	Yes	No	Total
Yes	38%	16%	23%
No	62	84	77
Total	100	100	100

DESSERT ORDERED	BEEF ENTRÉE		
	Yes	No	Total
Yes	11.75%	10.79%	22.54%
No	19.52	57.94	77.46
Total	31.27	68.73	100

(b) If the owner is interested in finding out the percentage of males and females who order dessert or the percentage of those who order a beef entrée and a dessert among all patrons, the table of total percentages is most informative. If the owner is interested in the effect of gender on ordering of dessert or the effect of ordering a beef entrée on the ordering of dessert, the table of column percentages will be most informative. Because dessert is usually ordered after the main entrée, and the owner has no direct control over the gender of patrons, the table of row percentages is not very useful here. **(c)** 17% of the men ordered desserts, compared to 29% of the women; women are almost twice as likely to order dessert as men. Almost 38% of the patrons ordering a beef entrée ordered dessert, compared to 16% of patrons ordering all other entrées. Patrons ordering beef are more than 2.3 times as likely to order dessert as patrons ordering any other entrée.

2.86 (a) Most of the complaints were against the airlines. **(c)** Most of the complaints against U.S. airlines were about flight problems, followed by reservations/ticketing/boarding, customer service, and baggage. **(d)** Most of the complaints against foreign airlines were about baggage, then reservations/ticketing/boarding, flight problems, and customer service.

2.88 (c) The alcohol percentage is concentrated between 4% and 6%, with more between 4% and 5%. The calories are concentrated between 140 and 160. The carbohydrates are concentrated between 12 and 15. There are outliers in the percentage of alcohol in both tails. The outlier in the lower tail is due to the non-alcoholic beer O'Doul's, with only a 0.4% alcohol content. There are a few beers with alcohol content as high as around 11.5%. There are a few beers with calorie content as high as around 327.5 and carbohydrates as high as 31.5. There is a strong positive relationship between percentage of alcohol and calories and between calories and carbohydrates, and there is a moderately positive relationship between percentage alcohol and carbohydrates.

2.90 (c) There appears to be a positive relationship between the yield of the one-year CD and the five-year CD.

2.92 (a)**Frequency (Boston)**

Weight (Boston)	Frequency	Percentage
3,015 but less than 3,050	2	0.54%
3,050 but less than 3,085	44	11.96
3,085 but less than 3,120	122	33.15
3,120 but less than 3,155	131	35.60
3,155 but less than 3,190	58	15.76
3,190 but less than 3,225	7	1.90
3,225 but less than 3,260	3	0.82
3,260 but less than 3,295	1	0.27

(b)**Frequency (Vermont)**

Weight (Vermont)	Frequency	Percentage
3,550 but less than 3,600	4	1.21%
3,600 but less than 3,650	31	9.39
3,650 but less than 3,700	115	34.85
3,700 but less than 3,750	131	39.70
3,750 but less than 3,800	36	10.91
3,800 but less than 3,850	12	3.64
3,850 but less than 3,900	1	0.30

(d) 0.54% of the Boston shingles pallets are underweight and 0.27% are overweight. 1.21% of the Vermont shingles pallets are underweight and 3.94% are overweight.

2.94 (c)

Calories	Frequency	Percentage	Percentage	
			Limit	Less Than
50 but less than 100	3	12%	100	12%
100 but less than 150	3	12	150	24
150 but less than 200	9	36	200	60
200 but less than 250	6	24	250	84
250 but less than 300	3	12	300	96
300 but less than 350	0	0	350	96
350 but less than 400	1	4	400	100

Cholesterol	Frequency	Percentage	Percentage	
			Limit	Less Than
0 but less than 50	2	8%	50	8%
50 but less than 100	17	68	100	76
100 but less than 150	4	16	150	92
150 but less than 200	1	4	200	96
200 but less than 250	0	0	250	96
250 but less than 300	0	0	300	96
300 but less than 350	0	0	350	96
350 but less than 400	0	0	400	96
400 but less than 450	0	0	450	96
450 but less than 500	1	4	500	100

The sampled fresh red meats, poultry, and fish vary from 98 to 397 calories per serving, with the highest concentration between 150 and 200 calories. One protein source, spareribs, with 397 calories, is more than 100 calories above the next-highest-caloric food. The protein content of the sampled foods varies from 16 to 33 grams, with 68% of the values falling between 24 and 32 grams. Spareribs and fried liver are both very different from other foods sampled—the former on calories and the latter on cholesterol content.

2.96 (b) There is a downward trend in the amount filled. **(c)** The amount filled in the next bottle will most likely be below 1.894 liters. **(d)** The scatter plot of the amount of soft drink filled against time reveals the trend of the data, whereas a histogram only provides information on the distribution of the data.

CHAPTER 3

3.2 (a) Mean = 7, median = 7, mode = 7. **(b)** Range = 9, $S^2 = 10.8$, $S = 3.286$, $CV = 46.948\%$. **(c)** Z scores: 0, -0.913, 0.609, 0, -1.217, 1.522. None of the Z scores are larger than 3.0 or smaller than -3.0. There is no outlier. **(d)** Symmetric because mean = median.

3.4 (a) Mean = 2, median = 7, mode = 7. **(b)** Range = 17, $S^2 = 62$, $S = 7.874$, $CV = 393.7\%$. **(c)** 0.635, -0.889, -1.270, 0.635, 0.889. There are no outliers. **(d)** Left-skewed because mean < median.

3.6 -0.0835.**3.8 (a)**

	Grade X	Grade Y
Mean	575	575.4
Median	575	575
Standard deviation	6.40	2.07

(b) If quality is measured by central tendency, Grade X tires provide slightly better quality because X's mean and median are both equal to the expected value, 575 mm. If, however, quality is measured by consistency, Grade Y provides better quality because, even though Y's mean is only slightly larger than the mean for Grade X, Y's standard deviation is much smaller. The range in values for Grade Y is 5 mm compared to the range in values for Grade X, which is 16 mm.

(c)

	Grade X	Grade Y, Altered
Mean	575	577.4
Median	575	575
Standard deviation	6.40	6.11

When the fifth Y tire measures 588 mm rather than 578 mm, Y's mean inner diameter becomes 577.4 mm, which is larger than X's mean inner diameter, and Y's standard deviation increases from 2.07 mm to 6.11 mm. In this case, X's tires are providing better quality in terms of the mean inner diameter, with only slightly more variation among the tires than Y's.

3.10 (a), (b)

	Cost (\$)
Mean	7.0933
Standard Error	0.3630
Median	6.8
Mode	6.5
Standard Deviation	1.4060
Sample Variance	1.9769
Kurtosis	-0.5778
Skewness	0.4403
Range	4.71
Minimum	4.89
Maximum	9.6
Sum	106.4
Count	15
First Quartile	5.9
Third Quartile	8.3
CV	19.82%

(c) The mean is only slightly greater than the median, so the data are only slightly right-skewed. (d) The mean amount spent is \$7.09, and the median is \$6.80. The average scatter of the amount spent around the mean is \$1.41. The difference between the highest and the lowest amount spent is \$4.71.

3.12 (a), (b)

MPG

MPG	
Mean	22.5294
Standard Error	0.4465
Median	22
Mode	22
Standard Deviation	1.8412
Sample Variance	3.3897
Kurtosis	0.3402
Skewness	0.5259
Range	7
Minimum	19
Maximum	26
Count	17
CV	8.17%

MPG	Z Score	MPG	Z Score
22	-0.2875	19	-1.9170
23	0.2556	22	-0.2875
21	-0.8307	22	0.2875
22	-0.2875	26	1.8850
25	1.3419	23	0.2556
26	1.8850	24	0.7987
22	-0.2875	21	-0.8307
22	-0.2875	22	-0.2875
21	-0.8307		

(c) Because the mean is slightly more than the median, the data are slightly right skewed. (d) The distribution of MPG of the sedans is right-skewed as is the distribution of the SUVs. The mean MPG of sedans is 4.59 higher than that of SUVs. The average scatter of the MPG of sedans is almost 3 times that of SUVs. The range of sedans is slightly more than 2.5 times that of SUVs.

3.14 (a), (b)

Facebook Penetration

Facebook Penetration	
Mean	36.2773
Standard Error	3.4074
Median	38.16
Mode	#N/A
Standard Deviation	13.1969
Sample Variance	174.1588
Kurtosis	0.7454
Skewness	-0.7218
Range	48.08
Minimum	5.37
Maximum	53.45
Count	15
CV	36.38%

Country	Facebook Penetration	Z Score
United States	52.56	1.2338
Brazil	33.09	-0.2415
India	5.37	-2.3420
Indonesia	19.41	-1.2781
Mexico	32.52	-0.2847
United Kingdom	51.61	1.1618
Turkey	41.69	0.4101
Philippines	30.12	-0.4666
France	39.07	0.2116
Germany	30.62	-0.4287
Italy	38.16	0.1427
Argentina	49.35	0.9906
Canada	53.45	1.3013
Colombia	40.01	0.2828
Thailand	27.13	-0.6931

None of the Z scores are more than three standard deviations away from the mean, so there are no outliers. (c) The mean is slightly smaller than the median, so the data are only slightly left-skewed. (d) The mean market penetration value is 36.2773 and the median is 38.16. The average scatter around the mean is 13.1969. The difference between the highest value and the lowest value is 48.08.

3.16 (a), (b)

Cost(English pounds)

Mean	155.75
Standard Error	4.566
Median	158
Mode	158
Standard Deviation	12.9146
Sample Variance	166.7857
Kurtosis	-0.9140
Skewness	-0.5196
Range	36
Minimum	135
Maximum	171
Count	8

(c) The mean price is \$155.5 and the median is \$158.50. The average scatter around the mean is \$12.91. The difference between the highest and the lowest value is \$36.

(d) (a), (b)

Cost(U.S. \$)

Mean	159.375
Standard Error	7.0557
Median	158
Mode	158
Standard Deviation	19.9566
Sample Variance	398.2679
Kurtosis	2.0333
Skewness	1.1149
Range	65
Minimum	135
Maximum	200
Count	8

(c) (d) The mean price is \$159.38 and the median is \$158.00. The average scatter around the mean is \$19.96. The difference between the highest and the lowest value is \$65. The mean, standard deviation, and range are sensitive to outliers. The higher price at \$200 raises the value of mean, standard deviation, and range but has no impact on the median. It also changed the skewness from negative to positive and increased the kurtosis statistic.

3.18 (a) Mean = 7.11, median = 6.68. **(b)** Variance = 4.336, standard deviation = 2.082, range = 6.67, CV = 29.27%. **(c)** Because the mean is greater than the median, the distribution is right-skewed. **(d)** The mean and median are both greater than five minutes. The distribution is right-skewed, meaning that there are some unusually high values. Further, 13 of the 15 bank customers sampled (or 86.7%) had waiting times greater than five minutes. So the customer is likely to experience a waiting time in excess of five minutes. The manager overstated the bank's service record in responding that the customer would "almost certainly" not wait longer than five minutes for service.

3.20 (a) $[(1 + 0.0240) \times (1 + 0.7460)]^{1/2} - 1 = 1.3371 - 1 = 33.71\%$ per year. **(b)** $= (\$1,000) \times (1 + 0.3371) \times (1 + 0.3371) = \$1,787.84$. **(c)** The result for Taser was better than the result for GE, which was worth \$1,188.41.

3.22 (a) Platinum = 12.90%, gold = 15.41%, silver = 27.58% per year. **(b)** Silver had a much higher return than gold or platinum. **(c)** Silver had a much higher return than the DJIA, the S&P 500, and the NASDAQ; gold's return was worse than the NASDAQ but better than the S&P 500 and DJIA; platinum's return was better than S&P 500 and DJIA but worse than NASDAQ.

3.24 (a)

Average of 1YrReturn% Star Rating						
Type	Five	Four	One	Three	Two	Grand Total
Growth	16.5544	15.2193	10.3575	13.9957	13.6058	14.2780
Large	18.0756	15.4971	12.3320	14.8743	17.1257	15.6771
Mid-Cap	15.5200	15.0400	10.0875	13.4140	8.7046	13.2160
Small	13.3300	15.0082	9.1014	12.7676	12.4722	12.9771
Value	17.2820	12.7295	13.4957	15.3603	15.4863	14.6982
Large	16.4150	11.7515	12.1120	14.5648	14.1633	13.5898
Mid-Cap	16.4400	16.1625		16.7267	17.4680	16.7879
Small	18.5700	12.5260	16.9550	16.0950	15.8860	15.4840
Grand Total	16.7126	14.6604	11.3126	14.4423	14.1821	14.3963

(b)

StdDev of 1YrReturn% Star Rating						
Type	Five	Four	One	Three	Two	Grand Total
Growth	4.0813	3.6946	5.0187	3.8308	7.6709	5.0041
Large	4.3119	4.1374	4.6690	2.7064	7.4925	4.7615
Mid-Cap	3.3099	3.1017	8.7458	4.8023	7.6199	5.4705
Small	4.4265	3.9244	2.2479	4.3906	2.9127	4.0854
Value	6.9822	4.5679	4.3343	4.1815	3.6530	4.4651
Large	1.1384	3.6990	4.3732	4.5739	2.5803	4.0592
Mid-Cap	#DIV/0!	4.1910		3.1676	4.5127	3.4837
Small	13.7179	6.3546	1.6476	3.9994	4.1620	5.4861
Grand Total	4.6722	4.0202	4.9474	3.9820	6.7243	4.8551

(c) The mean one-year return of small-cap value funds is higher than that of the small-cap growth funds across the different star ratings with the exception of those rated as four-star. On the other hand, the mean one-year return of large-cap value funds is lower than that of the growth funds across the different star ratings, but the mid-cap value funds are higher across the different star ratings.

The standard deviation of the one-year return of growth funds is generally higher than that of the value funds across all the star ratings and market caps with the exception of the large-cap and three-star, mid-cap and five-star, mid-cap and four-star, mid-cap and one-star, small-cap and five-star, small-cap and four-star, and small-cap and two-star.

3.26 (a)

Average of 1YrReturn% Star Rating						
Type	Five	Four	One	Three	Two	Grand Total
Growth	16.5544	15.2193	10.3575	13.9957	13.6058	14.2780
Average	16.5333	16.2233	11.6467	13.0514	10.8005	13.0524
High		14.6100	9.3620	14.5900	33.7200	17.7170
Low	16.5587	14.9785	9.8060	14.5700	13.6822	14.6717
Value	17.2820	12.7295	13.4957	15.3603	15.4863	14.6982
Average	28.2700		13.9800	16.4786	17.5267	17.1012
High			12.0500		22.1400	15.4133
Low	14.5350	12.7295	14.2150	15.0903	13.9117	14.0751
Grand Total	16.7126	14.6604	11.3126	14.4423	14.1821	14.3963

(b)

StdDev of 1YrReturn% Star Rating						
Type	Five	Four	One	Three	Two	Grand Total
Growth	4.0813	3.6946	5.0187	3.8308	7.6709	5.0041
Average	3.0735	4.9524	7.6948	4.9654	6.9272	6.0163
High		#DIV/0!	2.6945	#DIV/0!	0.2946	11.3821
Low	4.3448	3.3483	3.0114	2.8818	2.1220	3.3562
Value	6.9822	4.5679	4.3343	4.1815	3.6530	4.4651
Average	#DIV/0!		4.0506	2.9673	3.8277	4.4488
High			8.5843		#DIV/0!	8.4131
Low	3.8335	4.5679	0.5445	4.4251	2.4852	4.1475
Grand Total	4.6722	4.0202	4.9474	3.9820	6.7243	4.8551

(c) In general, the mean one-year return of the five-star rated growth funds is highest, followed by that of the four-star, three-star, two-star, and one-star rated growth funds across the various risk levels. However, a similar pattern does not hold through among the value funds.

There is no obvious pattern in the standard deviation of the one-year return.

3.28 (a) 4, 9, 5. **(b)** 3, 4, 7, 9, 12. **(c)** The distances between the median and the extremes are close, 4 and 5, but the differences in the tails are different (1 on the left and 3 on the right), so this distribution is slightly right-skewed. **(d)** In Problem 3.2 (d), because mean = median, the distribution is symmetric. The box part of the graph is symmetric, but the tails show right-skewness.

3.30 (a) -6.5, 8, 14.5. **(b)** -8, -6.5, 7, 8, 9. **(c)** The shape is left-skewed. **(d)** This is consistent with the answer in Problem 3.4 (d).

3.32 (a), (b)

Five-Number Summary

Minimum	5.37
First Quartile	30.12
Median	38.16
Third Quartile	49.35
Maximum	53.45
Interquartile Range	19.23

The penetration value is left-skewed.

3.34 (a), (b)**Five-Number Summary**

Minimum	19
First Quartile	21.5
Median	22
Third Quartile	23.5
Maximum	26
Interquartile Range	2

(c) The MPG is right-skewed.

3.36 (a) Commercial district five-number summary: 0.38 3.2 4.5 5.55 6.46. Residential area five-number summary: 3.82 5.64 6.68 8.73 10.49. **(b)** Commercial district: The distribution is left-skewed. Residential area: The distribution is slightly right-skewed. **(c)** The central tendency of the waiting times for the bank branch located in the commercial district of a city is lower than that of the branch located in the residential area. There are a few long waiting times for the branch located in the residential area, whereas there are a few exceptionally short waiting times for the branch located in the commercial area.

3.38 (a) Population mean, $\mu = 6$. **(b)** Population standard deviation, $\sigma = 1.673$, population variance, $\sigma^2 = 2.8$.

3.40 (a) 68%. **(b)** 95%. **(c)** Not calculable, 75%, 88.89%. **(d)** $\mu - 4\sigma$ to $\mu + 4\sigma$ or -2.8 to 19.2 .

3.42 (a)

$$\text{Mean} = \frac{662,960}{51} = 12,999.22, \text{ variance} = \frac{762,944,726.6}{51} = 14,959,700.52,$$

standard deviation = $\sqrt{14,959,700.52} = 3,867.78$. **(b)** 64.71%, 98.04%, and 100% of these states have mean per capita energy consumption within 1, 2, and 3 standard deviations of the mean, respectively. **(c)** This is consistent with 68%, 95%, and 99.7%,

according to the empirical rule. **(d)** **(a)** Mean = $\frac{642,887}{50} = 12,857.74$,

$$\text{variance} = \frac{711,905,533.6}{50} = 14,238,110.67, \text{ standard deviation}$$

= $\sqrt{14,238,110.67} = 3,773.34$. **(b)** 66%, 98%, and 100% of these states have a mean per capita energy consumption within 1, 2, and 3 standard deviations of the mean, respectively. **(c)** This is consistent with 68%, 95%, and 99.7%, according to the empirical rule.

3.44 Covariance = 65.2909, $r = +1.0$.

$$\text{3.46 (a)} \text{ cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{800}{6} = 133.3333.$$

$$\text{(b)} r = \frac{\text{cov}(X, Y)}{S_X S_Y} = \frac{133.3333}{(46.9042)(3.3877)} = 0.8391.$$

(c) The correlation coefficient is more valuable for expressing the relationship between calories and sugar because it does not depend on the units used to measure calories and sugar. **(d)** There is a strong positive linear relationship between calories and sugar.

3.48 (a) $\text{cov}(X, Y) = 1.4115 \times 10^{13}$ **(b)** $r = 0.7752$ **(c)** There is a positive linear relationship between the coaches' total pay and revenue.

3.64 (a) Mean = 43.89, median = 45, 1st quartile = 18, 3rd quartile = 63. **(b)** Range = 76, interquartile range = 45, variance = 639.2564, standard deviation = 25.28, $CV = 57.61\%$. **(c)** The distribution is right-skewed because there are a few policies that require an exceptionally long period to be approved. **(d)** The mean approval process takes 43.89 days, with 50% of the policies being approved in less than 45 days. 50% of the applications are approved between 18 and 63 days. About 67% of the applications are approved between 18.6 and 69.2 days.

3.66 (a) Mean = 8.421, median = 8.42, range = 0.186, $S = 0.0461$. The mean and median width are both 8.42 inches. The range of the widths is 0.186 inch, and the average scatter around the mean is 0.0461 inch. **(b)** 8.312, 8.404, 8.42, 8.459, 8.498. **(c)** Even though the mean = median, the left tail is slightly longer, so the distribution is slightly left-skewed. **(d)** All the troughs in this sample meet the specifications.

3.68 (a), (b)

	Bundle Score	Typical Cost (\$)
Mean	54.775	24.175
Standard Error	4.3673	2.8662
Median	62	20
Mode	75	8
Standard Deviation	27.6215	328.6096
Sample Variance	762.9481	18.1276
Kurtosis	-0.8454	2.7664
Skewness	-0.4804	1.5412
Range	98	83
Minimum	2	5
Maximum	100	88
Sum	2,191	967
Count	40	40
First Quartile	34	9
Third Quartile	75	31
Interquartile Range	41	22
CV	50.43%	74.98%

(c) The typical cost is right-skewed, while the bundle score is left-skewed.

(d) $r = 0.3465$. **(e)** The mean typical cost is \$24.18, with an average spread around the mean equaling \$18.13. The spread between the lowest and highest costs is \$83. The middle 50% of the typical cost fall over a range of \$22 from \$9 to \$31, while half of the typical cost is below \$20. The mean bundle score is 54.775, with an average spread around the mean equaling 27.6215. The spread between the lowest and highest scores is 98. The middle 50% of the scores fall over a range of 41 from 34 to 75, while half of the scores are below 62. The typical cost is right-skewed, while the bundle score is left-skewed. There is a weak positive linear relationship between typical cost and bundle score.

3.70 (a) Boston: 0.04, 0.17, 0.23, 0.32, 0.98; Vermont: 0.02, 0.13, 0.20, 0.28, 0.83. **(b)** Both distributions are right-skewed. **(c)** Both sets of shingles did quite well in achieving a granule loss of 0.8 gram or less. Only two Boston shingles had a granule loss greater than 0.8 gram. The next highest to these was 0.6 gram. These two values can be considered outliers. Only 1.176% of the shingles failed the specification. Only one of the Vermont shingles had a granule loss greater than 0.8 gram. The next highest was 0.58 gram. Thus, only 0.714% of the shingles failed to meet the specification.

3.72 (a) The correlation between calories and protein is 0.4644. **(b)** The correlation between calories and cholesterol is 0.1777. **(c)** The correlation between protein and cholesterol is 0.1417. **(d)** There is a weak positive linear relationship between calories and protein, with a correlation coefficient of 0.46. The positive linear relationships between calories and cholesterol and between protein and cholesterol are very weak.

3.74 (a), (b)

Property Taxes per Capita (\$)

Mean	1,332.2353
Median	1,230
Standard deviation	577.8308
Sample variance	333,888.4235
Range	2,479
First quartile	867
Third quartile	1,633
Interquartile range	766
Coefficient of variation	43.37%

(c), (d) The distribution of the property taxes per capita is right-skewed, with a mean value of \$1,332.24, a median of \$1,230, and an average spread around the mean of \$577.83. There is an outlier in the right tail at \$2,985, while the standard deviation is about 43.37% of the mean. 25% of the states have property tax that falls below \$867 per capita, and 25% have property taxes that are higher than \$1,633 per capita.

3.76 (a), (b)

	Abandonment rate in % (7:00AM–3:00PM)
Mean	13.8636
Standard Error	1.625414306
Median	10
Mode	9
Standard Deviation	7.6238688759
Sample Variance	58.1233
Kurtosis	0.7235687396
Skewness	1.1807
Range	29
Minimum	5
Maximum	34
Sum	305
Count	22
First Quartile	9
Third Quartile	20
Interquartile Range	11
CV	54.99%

(c) The data are right-skewed.

(d) $r = 0.7575$

(e) The mean abandonment rate is 13.86%. Half of the abandonment rates are less than 10%. One-quarter of the abandonment rates are less than 9% while another one-quarter are more than 20%. The overall spread of the abandonment rates is 29%. The middle 50% of the abandonment rates are spread over 11%. The average spread of abandonment rates around the mean is 7.62%. The abandonment rates are right-skewed.

3.78 (a), (b)

Average Credit Score	
Mean	746.2238
Standard Error	1.821396
Median	749
Mode	760
Standard Deviation	21.78073
Sample Variance	474.4003
Kurtosis	-0.83035
Skewness	-0.22982
Range	89
Minimum	700
Maximum	789
Sum	106710
Count	143
First Quartile	730
Third Quartile	763
Interquartile Range	33
CV	2.92%

(c) The data are symmetrical.

(d) The mean of the average credit scores is 746.2238. Half of the average credit scores are less than 749. One-quarter of the average credit scores are less than 730 while another one-quarter is more than 763. The overall spread of average credit scores is 89. The middle 50% of the average credit scores spread over 33. The average spread of average credit scores around the mean is 21.7807.

CHAPTER 4

4.2 (a) Simple events include selecting a red ball. **(b)** Selecting a white ball. **(c)** The sample space consists of the 12 red balls and the 8 white balls.

4.4 (a) $60/100 = 3/5 = 0.6$. **(b)** $10/100 = 1/10 = 0.1$.
(c) $35/100 = 7/20 = 0.35$. **(d)** $9/10 = 0.9$.

4.6 (a) Mutually exclusive, not collectively exhaustive. **(b)** Not mutually exclusive, not collectively exhaustive. **(c)** Mutually exclusive, not collectively exhaustive. **(d)** Mutually exclusive, collectively exhaustive.

4.8 (a) Is a male.
(b) Is a male and feels tense or stressed out at work.
(c) Does not feel tense or stressed out at work.
(d) Is a male and feels tense or stressed out at work is a joint event because it consists of two characteristics.

4.10 (a) A marketer who plans to increase use of LinkedIn. **(b)** A B2B marketer who plans to increase use of LinkedIn. **(c)** A marketer who does not plan to increase use of LinkedIn. **(d)** A marketer who plans to increase use of LinkedIn and is a B2C marketer is a joint event because it consists of two characteristics, plans to increase use of LinkedIn and is a B2C marketer.

4.12 (a) $8,007/14,074 = 0.5689$. **(b)** $6,264/14,074 = 0.4451$.
(c) $8,007/14,074 + 6,264/14,074 - 3,633/14,074 = 0.7559$
(d) The probability of saying that analyzing data is critical or is a manager includes the probability of saying that analyzing data is critical plus the probability of being a manager minus the joint probability of saying that analyzing data is critical and is a manager.

4.14 (a) 514/1,085. (b) 276/1,085. (c) 781/1,085.
 (d) $1,085/1,085 = 1.00$.

4.16 (a) $10/30 = 1/3 = 0.33$. (b) $20/60 = 1/3 = 0.33$.
 (c) $40/60 = 2/3 = 0.67$. (d) Because $P(A|B) = P(A) = 1/3$, events A and B are independent.

4.18 $\frac{1}{2} = 0.5$.

4.20 Because $P(A \text{ and } B) = 0.20$ and $P(A)P(B) = 0.12$, events A and B are not independent.

4.22 (a) $1,478/1,945 = 0.7599$. (b) $1,027/1,868 = 0.5498$.
 (c) $P(\text{Increased use of LinkedIn}) = 2,505/3,813 = 0.6570$, which is not equal to $P(\text{Increased use of LinkedIn} | \text{B2B}) = 0.7599$. Therefore, increased use of LinkedIn and business focus are not independent.

4.24 (a) $4,374/7,810 = 0.5601$. (b) $3,436/7,810 = 0.4399$.
 (c) $3,633/6,264 = 0.5800$. (d) $2,631/6,264 = 0.4200$.

4.26 (a) $0.025/0.6 = 0.0417$. (b) $0.015/0.4 = 0.0375$. (c) Because $P(\text{Needs warranty repair} | \text{Manufacturer based in U.S.}) = 0.0417$ and $P(\text{Needs warranty repair}) = 0.04$, the two events are not independent.

4.28 (a) 0.0045. (b) 0.012. (c) 0.0059. (d) 0.0483.

4.30 0.095.

4.32 (a) 0.736. (b) 0.997.

4.34 (a) $P(B' | O) = \frac{(0.5)(0.3)}{(0.5)(0.3) + (0.25)(0.7)} = 0.4615$.
 (b) $P(O) = 0.175 + 0.15 = 0.325$.

4.36 (a) $P(\text{Huge success} | \text{Favorable review}) = 0.099/0.459 = 0.2157$;
 $P(\text{Moderate success} | \text{Favorable review}) = 0.14/0.459 = 0.3050$;
 $P(\text{Break even} | \text{Favorable review}) = 0.16/0.459 = 0.3486$;
 $P(\text{Loser} | \text{Favorable review}) = 0.06/0.459 = 0.1307$.
 (b) $P(\text{Favorable review}) = 0.459$.

4.38 $3^{10} = 59,049$.

4.40 (a) $2^7 = 128$. (b) $6^7 = 279,936$. (c) There are two mutually exclusive and collectively exhaustive outcomes in (a) and six in (b).

4.42 $(8)(4)(3)(3) = 288$.

4.44 $5! = (5)(4)(3)(2)(1) = 120$. Not all the orders are equally likely because the teams have a different probability of finishing first through fifth.

4.46 $6! = 720$.

4.48 $\frac{10!}{4!6!} = 210$.

4.50 4,950.

4.60 (a)

SHARE HEALTH INFORMATION	AGE		
	18–24	45–64	Total
Yes	400	225	625
No	100	275	375
Total	500	500	1,000

(b) Simple event: “Shares health information through social media.”
 Joint event: “Shares health information through social media and

is between 18 and 24 years old.” (c) $P(\text{Shares health information through social media}) = 675/1,000 = 0.675$. (d) $P(\text{Shares health information through social media and is in the 45-to-64-year-old group}) = 225/1000 = 0.225$. (e) Not independent.

4.62 (a) 84/200. (b) 126/200. (c) 141/200. (d) 33/200. (f) 16/100.

4.64 (a) $202/447 = 0.4519$. (b) $95/237 = 0.4008$. (c) $107/210 = 0.5095$. (d) $217/447 = 0.4855$. (e) $122/237 = 0.5148$. (f) $95/210 = 0.4524$. (g) IT executives were more likely to identify big data as critical while marketing executives were more likely to identify functional silos as an issue.

CHAPTER 5

5.2 (a)

$$\mu = 0(0.10) + 1(0.20) + 2(0.45) + 3(0.15) + 4(0.05) + 5(0.05) = 2.0.$$

$$(b) \sigma = \sqrt{\frac{(0-2)^2(0.10) + (1-2)^2(0.20) + (2-2)^2(0.45) + (3-2)^2(0.15) + (4-2)^2(0.05) + (5-2)^2(0.05)}{(3-2)^2(0.15) + (4-2)^2(0.05) + (5-2)^2(0.05)}} = 1.183.$$

	X	P(X)
\$ - 1		21/36
\$ + 1		15/36

	X	P(X)
\$ - 1		21/36
\$ + 1		15/36

	X	P(X)
\$ - 1		30/36
\$ + 1		6/36

(d) $-\$0.167$ for each method of play.

5.6 (a) 2.1058. (b) 1.4671.

5.8 (a) 90; 30. (b) 126.10, 10.95. (c) $-1,300$. (d) 120.

5.10 (a) 9.5 minutes. (b) 1.9209 minutes.

5.12

X × P(X)	Y × P(Y)	$(X - \mu_X)^2$	$(Y - \mu_Y)^2$	$(Y - \mu_Y) \times P(x, y)$
-10	5	2,528.1	129.6	-572.4
0	45	1,044.3	5,548.8	-2,407.2
24	-6	132.3	346.8	-214.2
45	-30	2,484.3	3,898.8	-3,112.2

$$(a) E(X) = \mu_X = \sum_{i=1}^N X_i P(X_i) = 59, E(Y) = \mu_Y = \sum_{i=1}^N Y_i P(Y_i) = 14.$$

$$(b) \sigma_X = \sqrt{\sum_{i=1}^N [X_i - E(X)]^2 P(X_i)} = 78.6702.$$

$$\sigma_Y = \sqrt{\sum_{i=1}^N [Y_i - E(Y)]^2 P(Y_i)} = 99.62.$$

$$(c) \sigma_{XY} = \sum_{i=1}^N [X_i - E(X)][Y_i - E(Y)] P(x_i, y_i) = -6,306.$$

(d) Stock X gives the investor a lower standard deviation while yielding a higher expected return, so the investor should select stock X.

5.14 (a) \$71; \$97. **(b)** 61.88; 84.27. **(c)** 5,113. **(d)** Risk-averse investors would invest in stock X, whereas risk takers would invest in stock Y.

5.16 (a) $E(X) = \$66.20$; $E(Y) = \$63.01$. **(b)** $\sigma_X = \$57.22$; $\sigma_Y = \$195.22$. **(c)** $\sigma_{XY} = \$10,766.44$. **(d)** Based on the expected value criteria, you would choose the common stock fund. However, the common stock fund also has a standard deviation more than three times higher than that for the corporate bond fund. An investor should carefully weigh the increased risk. **(e)** If you chose the common stock fund, you would need to assess your reaction to the small possibility that you could lose virtually all of your entire investment.

5.18 (a) 0.5997. **(b)** 0.0016. **(c)** 0.0439. **(d)** 0.4018.

5.20 (a) 0.40, 0.60. **(b)** 1.60, 0.98. **(c)** 4.0, 0.894. **(d)** 1.50, 0.866.

5.22 (a) 0.0425. **(b)** 0.0004. **(c)** 0.0492. **(d)** $\mu = 1.62$, $\sigma = 1.0875$. **(e)** Each under 25 year old either owns a tablet or does not own a tablet and each person surveyed is independent of every other person.

5.24 (a) 0.5987. **(b)** 0.3151. **(c)** 0.9885. **(d)** 0.0115.

5.26 (a) 0.5718. **(b)** 0.0049. **(c)** 0.9231. **(d)** $\mu = 2.49$, $\sigma = 0.6506$.

5.28 (a) 0.2565. **(b)** 0.1396. **(c)** 0.3033. **(d)** 0.0247.

5.30 (a) 0.0337. **(b)** 0.0067. **(c)** 0.9596. **(d)** 0.0404.

5.32 (a)

$$\begin{aligned} P(X < 5) &= P(X = 0) + P(X = 1) + P(x = 2) + P(X = 3) \\ &\quad + P(X = 4) \\ &= \frac{e^{-6}(6)^0}{0!} + \frac{e^{-6}(6)^1}{1!} + \frac{e^{-6}(6)^2}{2!} + \frac{e^{-6}(6)^3}{3!} + \frac{e^{-6}(6)^4}{4!} \\ &= 0.002479 + 0.014873 + 0.044618 + 0.089235 \\ &\quad + 0.133853 \\ &= 0.2851. \end{aligned}$$

$$\text{(b)} P(X = 5) = \frac{e^{-6}(6)^5}{5!} = 0.1606.$$

$$\text{(c)} P(X \geq 5) = 1 - P(X < 5) = 1 - 0.2851 = 0.7149.$$

$$\text{(d)} P(X = 4 \text{ or } X = 5) = P(X = 4) + P(X = 5) = \frac{e^{-6}(6)^4}{4!} + \frac{e^{-6}(6)^5}{5!} \\ = 0.2945.$$

5.34 (a) 0.1287. **(b)** 0.8713. **(c)** 0.6074.

5.36 (a) 0.1165. **(b)** 0.2504. **(c)** 0.6331. **(d)** 0.3669.

5.38 (a) 0.3263. **(b)** 0.8964. **(c)** Because Ford had a higher mean rate of problems per car than Toyota, the probability of a randomly selected Ford having zero problems and the probability of no more than two problems are both lower than for Toyota.

5.40 (a) 0.3535 **(b)** 0.9122. **(c)** Because Toyota had a lower mean rate of problems per car in 2009 compared to 2010, the probability of a randomly selected Toyota having zero problems and the probability of no more than two problems are both higher in 2009 than in 2010.

5.42 (a) 0.238. **(b)** 0.2. **(c)** 0.1591. **(d)** 0.0083.

5.44 (a) If $n = 6$, $E = 25$, and $N = 100$,

$$P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)]$$

$$= 1 - \left[\frac{\binom{25}{0} \binom{100-25}{6-0}}{\binom{100}{6}} + \frac{\binom{25}{1} \binom{100-25}{6-1}}{\binom{100}{6}} \right]$$

$$= 1 - [0.1689 + 0.3620] = 0.4691.$$

(b) If $n = 6$, $E = 30$, and $N = 100$,

$$P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)]$$

$$= 1 - \left[\frac{\binom{30}{0} \binom{100-30}{6-0}}{\binom{100}{6}} + \frac{\binom{30}{1} \binom{100-30}{6-1}}{\binom{100}{6}} \right]$$

$$= 1 - [0.1100 + 0.3046] = 0.5854.$$

(c) If $n = 6$, $E = 5$, and $N = 100$,

$$P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)]$$

$$= 1 - \left[\frac{\binom{5}{0} \binom{100-5}{6-0}}{\binom{100}{6}} + \frac{\binom{5}{1} \binom{100-5}{6-1}}{\binom{100}{6}} \right]$$

$$= 1 - [0.7291 + 0.2430] = 0.0279.$$

(d) If $n = 6$, $E = 10$, and $N = 100$,

$$P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)]$$

$$= 1 - \left[\frac{\binom{10}{0} \binom{100-10}{6-0}}{\binom{100}{6}} + \frac{\binom{10}{1} \binom{100-10}{6-1}}{\binom{100}{6}} \right]$$

$$= 1 - [0.5223 + 0.3687] = 0.1090.$$

(e) The probability that the entire group will be audited is very sensitive to the true number of improper returns in the population. If the true number is very low ($E = 5$), the probability is very low (0.0279). When the true number is increased by a factor of 6 ($E = 30$), the probability the group will be audited increases by a factor of more than 20 (0.5854).

5.46 (a) $P(X = 4) = 0.00003649$. **(b)** $P(X = 0) = 0.5455$.

(c) $P(X \geq 1) = 0.4545$. **(d)** $E = 6$. **(a)** $P(X = 4) = 0.0005$.

(b) $P(X = 0) = 0.3877$. **(c)** $P(X \geq 1) = 0.6123$.

5.48 (a) $P(X = 1) = 0.2424$. **(b)** $P(X \geq 1) = 0.9697$.

(c) $P(X = 3) = 0.2424$. **(d)** Because there were now 12 funds to consider, the probability that 3 would be growth funds decreased from 0.3810 to 0.2424.

5.54 (a) 0.65. **(b)** 0.65. **(c)** 0.3124. **(d)** 0.0053. **(e)** The assumption of independence may not be true.

5.56 (a) If $\pi = 0.50$ and $n = 13$, $P(X \geq 10) = 0.0461$.

(b) If $\pi = 0.75$ and $n = 13$, $P(X \geq 10) = 0.5843$.

5.58 (a) 0.0060. **(b)** 0.2007. **(c)** 0.1662. **(d)** Mean = 4.0, standard deviation = 1.5492. **(e)** Since the percentage of bills containing an error is lower in this problem, the probability is higher in (a) and (b) of this problem and lower in (c).

- 5.60** (a) $\mu = n\pi = 8.2$ (b) $\sigma = \sqrt{n\pi(1 - \rho)} = 2.1995$.
 (c) $P(X = 10) = 0.1268$. (d) $P(X \leq 5) = 0.1079$.
 (e) $P(X \geq 5) = 0.9577$.

5.62 (a) If $\pi = 0.50$ and $n = 40$, $P(X \geq 35) = 0.000000691$. (b) If $\pi = 0.70$ and $n = 40$, $P(X \geq 35) = 0.008618$. (c) If $\pi = 0.90$ and $n = 40$, $P(X \geq 35) = 0.793727$. (d) Based on the results in (a)–(c), the probability that the Standard & Poor's 500 Index will increase if there is an early gain in the first five trading days of the year is very likely to be close to 0.90 because that yields a probability of 79.37% that at least 35 of the 40 years the Standard & Poor's 500 Index will increase the entire year.

5.64 (a) The assumptions needed are (i) the probability that a questionable claim is referred by an investigator is constant, (ii) the probability that a questionable claim is referred by an investigator approaches 0 as the interval gets smaller, and (iii) the probability that a questionable claim is referred by an investigator is independent from interval to interval.
 (b) 0.0378. (c) 0.5830. (d) 0.4170.

CHAPTER 6

- 6.2** (a) 0.9089. (b) 0.0911. (c) +1.96. (d) -1.00 and +1.00.

- 6.4** (a) 0.1401. (b) 0.4168. (c) 0.3918. (d) +1.00.

- 6.6** (a) 0.9599. (b) 0.0228. (c) 43.42. (d) 46.64 and 53.36.

6.8 (a) $P(34 < X < 50) = P(-1.33 < Z < 0) = 0.4082$.
 (b) $P(X < 30) + P(X > 60) = P(Z < -1.67) + P(Z > 0.83) = 0.0475 + (1.0 - 0.7967) = 0.2508$. (c) $P(Z < -0.84) \cong 0.20$, $Z = -0.84 = \frac{X - 50}{12}$, $X = 50 - 0.84(12) = 39.92$ thousand miles, or 39,920 miles. (d) The smaller standard deviation makes the absolute Z values larger. (a) $P(34 < X < 50) = P(-1.60 < Z < 0) = 0.4452$. (b) $P(X < 30) + P(X > 60) = P(Z < -2.00) + P(Z > 1.00) = 0.0228 + (1.0 - 0.8413) = 0.1815$. (c) $X = 50 - 0.84(10) = 41.6$ thousand miles, or 41,600 miles.

6.10 (a) 0.9878. (b) 0.8185. (c) 86.16%. (d) Option 1: Because your score of 81% on this exam represents a Z score of 1.00, which is below the minimum Z score of 1.28, you will not earn an A grade on the exam under this grading option. Option 2: Because your score of 68% on this exam represents a Z score of 2.00, which is well above the minimum Z score of 1.28, you will earn an A grade on the exam under this grading option. You should prefer Option 2.

- 6.12** (a) 0.1265. (b) 0.1395. (c) 0.0192. (d) 76.5689.

6.14 With 39 values, the smallest of the standard normal quantile values covers an area under the normal curve of 0.025. The corresponding Z value is -1.96. The middle (20th) value has a cumulative area of 0.50 and a corresponding Z value of 0.0. The largest of the standard normal quantile values covers an area under the normal curve of 0.975, and its corresponding Z value is +1.96.

6.16 (a) Mean = 22.5294, median = 22, $S = 1.8411$, range = 7, $6S = 6(1.8411) = 11.0466$, interquartile range = 2.0, $1.33(1.8411) = 2.4487$. The mean is slightly more than the median. The range is much less than $6S$, and the interquartile range is less than $1.33S$. (b) The normal probability plot appears to be approximately normally distributed. The kurtosis is 0.3402, indicating very little departure from a normal distribution.

6.18 (a) Mean = 1,332.24, median = 1,230, range = 2,479, $6(S) = 3,466.98$, interquartile range = 766, $1.33(S) = 768.51$. Because the mean is greater than the median, the interquartile range is slightly less than 1.33 times the standard deviation, and the range is much smaller than 6 times the standard deviation, the data appear to deviate from the normal distribution. (b) The normal probability plot suggests that the data appear to be right-skewed. The kurtosis is 0.5395 indicating a distribution that is slightly more peaked than a normal distribution, with more values in the tails.

6.20 (a) Interquartile range = 0.0025, $S = 0.0017$, range = 0.008, $1.33(S) = 0.0023$, $6(S) = 0.0102$. Because the interquartile range is close to $1.33S$ and the range is also close to $6S$, the data appear to be approximately normally distributed. (b) The normal probability plot suggests that the data appear to be approximately normally distributed.

6.22 (a) Five-number summary: 82 127 148.5 168 213; mean = 147.06, mode = 130, range = 131, interquartile range = 41, standard deviation = 31.69. The mean is very close to the median. The five-number summary suggests that the distribution is approximately symmetric around the median. The interquartile range is very close to $1.33S$. The range is about \$50 below $6S$. In general, the distribution of the data appears to closely resemble a normal distribution. (b) The normal probability plot confirms that the data appear to be approximately normally distributed.

6.24 (a) $(20-0)/120 = 0.1667$. (b) $(30-10)/120 = 0.1667$. (c) $(120-35)/120 = 0.7083$. (d) Mean = 60, standard deviation = 34.641.

6.26 (a) 0.1667. (b) 0.1667. (c) 0.3333. (d) Mean = 21, standard deviation = 1.7321.

6.28 (a) 0.6321. (b) 0.3679. (c) 0.2326. (d) 0.7674.

6.30 (a) 0.7769. (b) 0.2231. (c) 0.1410. (d) 0.8590.

6.32 (a) For $\lambda = 2$, $P(X \leq 1) = 0.8647$. (b) For $\lambda = 2$, $P(X \leq 5) = 0.99996$. (c) For $\lambda = 1$, $P(X \leq 1) = 0.6321$, for $\lambda = 1$, $P(X \leq 5) = 0.9933$.

6.34 (a) 0.4512. (b) 0.3012. (c) 0.1816.

6.36 (a) 0.8647. (b) 0.3297. (c) (a) 0.9765. (b) 0.5276.

6.46 (a) 0.4772. (b) 0.9544. (c) 0.0456. (d) 1.8835. (e) 1.8710 and 2.1290.

6.48 (a) 0.1405. (b) 0.0256. (c) \$2,179.78. (d) \$898.22 to \$2,179.78.

6.50 (a) Waiting time will more closely resemble an exponential distribution. (b) Seating time will more closely resemble a normal distribution. (c) Both the histogram and normal probability plot suggest that waiting time more closely resembles an exponential distribution. (d) Both the histogram and normal probability plot suggest that seating time more closely resembles a normal distribution.

6.52 (a) 0.0426. (b) 0.0731. (c) 0.9696. (d) 1.2127. (e) 1.6891 to 6.7850. (f) 0.125, 0.125, 0.90.

CHAPTER 7

- 7.2** (a) Virtually 0. (b) 0.1587. (c) 0.0139. (d) 50.195.

7.4 (a) Both means are equal to 6. This property is called unbiasedness. (c) The distribution for $n = 3$ has less variability. The larger sample size has resulted in sample means being closer to μ .

7.6 (a) When $n = 4$, because the mean is larger than the median, the distribution of the sales price of new houses is skewed to the right, and so is the sampling distribution of \bar{X} although it will be less skewed than

the population. **(b)** If you select samples of $n = 100$, the shape of the sampling distribution of the sample mean will be very close to a normal distribution, with a mean of \$291,200 and a standard deviation of \$9,000. **(c)** 0.9959. **(d)** 0.3181.

7.8 (a) $P(\bar{X} > 26) = P(Z > -1.00) = 1.0 - 0.1587 = 0.8413.$

(b) $P(Z < 1.04) = 0.85$; $\bar{X} = 27 + 1.04(1.0) = 28.04$. **(c)** To be able to use the standardized normal distribution as an approximation for the area under the curve, you must assume that the population is approximately symmetrical. **(d)** $P(Z < 1.04) = 0.85$; $\bar{X} = 27 + 1.04(0.50) = 27.52$.

7.10 (a) 0.40. **(b)** 0.0704.

7.12 (a) $\pi = 0.501$, $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.501(1-0.501)}{100}} = 0.05$

$$P(p > 0.55) = P(Z > 0.98) = 1.0 - 0.8365 = 0.1635.$$

(b) $\pi = 0.60$, $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.6(1-0.6)}{100}} = 0.04899$

$$P(p > 0.55) = P(Z > -1.021) = 1.0 - 0.1539 = 0.8461.$$

(c) $\pi = 0.49$, $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.49(1-0.49)}{100}} = 0.05$

$$P(p > 0.55) = P(Z > 1.20) = 1.0 - 0.8849 = 0.1151.$$

(d) Increasing the sample size by a factor of 4 decreases the standard error by a factor of 2.

(a) $P(p > 0.55) = P(Z > 1.96) = 1.0 - 0.9750 = 0.0250$.

(b) $P(p > 0.55) = P(Z > -2.04) = 1.0 - 0.0207 = 0.9793$.

(c) $P(p > 0.55) = P(Z > 2.40) = 1.0 - 0.9918 = 0.0082$.

7.14 (a) 0.8944. **(b)** 0.7887. **(c)** 0.3085. **(d)** **(a)** 0.9938. **(b)** 0.9876. **(c)** 0.1587.

7.16 (a) 0.7661. **(b)** The probability is 90% that the sample percentage will be contained between 0.1085 to 0.1915. **(c)** The probability is 95% that the sample percentage will be contained between 0.1005 and 0.1995.

7.18 (a) 0.0326. **(b)** 0.0001. **(c)** Increasing the sample size by a factor of 4 decreases the standard error by a factor of 2. The sampling distribution of the proportion becomes more concentrated around the true proportion of 0.39 and, hence, the probability in (b) becomes smaller than that in (a).

7.24 (a) 0.4999. **(b)** 0.00009. **(c)** 0. **(d)** 0. **(e)** 0.7518.

7.26 (a) 0.8944. **(b)** 4.617; 4.783. **(c)** 4.641.

7.28 (a) 0.0002. **(b)** 0.0577. **(c)** 0.9423.

CHAPTER 8

8.2 $114.68 \leq \mu \leq 135.32$.

8.4 Yes, it is true because 5% of intervals will not include the population mean.

8.6 (a) You would compute the mean first because you need the mean to compute the standard deviation. If you had a sample, you would compute the sample mean. If you had the population mean, you would compute the population standard deviation. **(b)** If you have a sample, you are computing the sample standard deviation, not the population standard deviation needed in Equation (8.1). If you have a population and have computed the population mean and population standard deviation, you don't need a confidence interval estimate of the population mean because you already know the mean.

8.8 Equation (8.1) assumes that you know the population standard deviation. Because you are selecting a sample of 100 from the population, you are computing a sample standard deviation, not the population standard deviation.

8.10 (a) $\bar{X} \pm Z \cdot \frac{\sigma}{\sqrt{n}} = 7,500 \pm 1.96 \cdot \frac{1,000}{\sqrt{64}}$; $7,255 \leq \mu \leq 7,745$.

(b) No, since the confidence interval does not include 8,000 hours the manufacturer cannot support a claim that the bulbs have a mean of 8,000 hours. **(c)** No. Because σ is known and $n = 64$, from the Central Limit Theorem, you know that the sampling distribution of \bar{X} is approximately normal. **(d)** The confidence interval is narrower, based on a population standard deviation of 800 hours rather than the original standard deviation of 1,000 hours. $\bar{X} \pm Z \times \frac{\sigma}{\sqrt{n}} = 7,500 \pm 1.96 \times \frac{800}{\sqrt{64}}$,

$7,304 \leq \mu \leq 7,696$. No, since the confidence interval does not include 8,000 the manufacturer cannot support a claim that the bulbs have a mean life of 8,000 hours.

8.12 (a) 2.2622. **(b)** 3.2498. **(c)** 2.0395. **(d)** 1.9977. **(e)** 1.7531.

8.14 $-0.12 \leq \mu \leq 11.84$, $2.00 \leq \mu \leq 6.00$. The presence of the outlier increases the sample mean and greatly inflates the sample standard deviation.

8.16 (a) $75 \pm (2.0049)(9)/\sqrt{55}$; $72.57 \leq \mu \leq 77.43$ **(b)** You can be 95% confident that the population mean amount of one-time gift is between \$72.57 and \$77.43.

8.18 (a) $6.31 \leq \mu \leq 7.87$. **(b)** You can be 95% confident that the population mean amount spent for lunch at a fast-food restaurant is between \$6.31 and \$7.87.

8.20 (a) $21.58 \leq \mu \leq 23.48$. **(b)** You can be 95% confident that the population mean miles per gallon of 2013 small SUVs is between 21.58 and 23.48. **(c)** Because the 95% confidence interval for population mean miles per gallon of 2013 small SUVs does not overlap with that for the population mean miles per gallon of 2013 family sedans, you can conclude that the population mean miles per gallon of 2013 small SUVs is lower than that of 2013 family sedans.

8.22 (a) $31.12 \leq \mu \leq 54.96$. **(b)** The number of days is approximately normally distributed. **(c)** No, the outliers skew the data. **(d)** Because the sample size is fairly large, at $n = 50$, the use of the t distribution is appropriate.

8.24 (a) $28.97 \leq \mu \leq 43.59$. **(b)** That the population distribution is normally distributed. **(c)** Both the normal probability plot and the boxplot show that the distribution of the Facebook penetration is left-skewed, so with the small sample size, the validity of the confidence interval is in question.

8.26 $0.19 \leq \pi \leq 0.31$.

8.28 (a) $p = \frac{X}{n} = \frac{135}{500} = 0.27$, $p \pm Z \sqrt{\frac{p(1-p)}{n}} = 0.27 \pm 2.58 \sqrt{\frac{0.27(0.73)}{500}}$; $0.2189 \leq \pi \leq 0.3211$. **(b)** The manager in charge of promotional programs can infer that the proportion of households that would upgrade to an improved cellphone if it were made available at a substantially reduced cost is somewhere between 0.22 and 0.32, with 99% confidence.

8.30 (a) $0.4863 \leq \pi \leq 0.5737$. **(b)** No, you cannot because the interval estimate includes 0.50 (50%). **(c)** $0.5162 \leq \pi \leq 0.5438$. Yes, you can, because the interval is above 0.50 (50%). **(d)** The larger the sample size, the narrower the confidence interval, holding everything else constant.

8.32 (a) $0.3587 \leq \pi \leq 0.4018$. (b) $0.2017 \leq \pi \leq 0.2384$. (c) Many more people use their phone to keep themselves occupied during commercials or breaks than check something they were watching on television.

8.34 $n = 35$.

8.36 $n = 1,041$.

$$\text{(a)} n = \frac{Z^2 \sigma^2}{e^2} = \frac{(1.96)^2(400)^2}{50^2} = 245.86. \text{ Use } n = 246.$$

$$\text{(b)} n = \frac{Z^2 \sigma^2}{e^2} = \frac{(1.96)^2(400)^2}{25^2} = 983.41. \text{ Use } n = 984.$$

8.40 $n = 97$.

8.42 (a) $n = 74$. (b) $n = 43$.

8.44 (a) $n = 246$. (b) $n = 385$. (c) $n = 554$. (d) When there is more variability in the population, a larger sample is needed to accurately estimate the mean.

8.46 (a) $p = 0.18; 0.1365 \leq \pi \leq 0.2235$. (b) $p = 0.13; 0.0919 \leq \pi \leq 0.1681$. (c) $p = 0.09; 0.0576 \leq \pi \leq 0.1224$. (d) (a) $n = 1,418$. (b) $n = 1,087$. (c) $n = 787$.

8.48 (a) If you conducted a follow-up study to estimate the population proportion of individuals who say that banking on their mobile device is convenient, you would use $\pi = 0.77$ in the sample size formula because it is based on past information on the proportion. (b) $n = 756$.

8.54 (a) $p = 0.88; 0.8667 \leq \pi \leq 0.8936$

$$p = 0.58; 0.5597 \leq \pi \leq 0.6005$$

$$p = 0.61; 0.5897 \leq \pi \leq 0.6300$$

$$p = 0.18; 0.1643 \leq \pi \leq 0.1961$$

$$p = 0.18; 0.1643 \leq \pi \leq 0.1961$$

(b) Most adults have a cellphone. Many adults have a desktop computer or a laptop computer. Some adults have an e book reader or a tablet computer.

8.56 (a) $39.88 \leq \mu \leq 42.12$. (b) $0.6158 \leq \pi \leq 0.8842$. (c) $n = 25$.

(d) $n = 267$. (e) If a single sample were to be selected for both purposes, the larger of the two sample sizes ($n = 267$) should be used.

8.58 (a) $3.19 \leq \mu \leq 9.21$. (b) $0.3242 \leq \pi \leq 0.7158$. (c) $n = 110$.

(d) $n = 121$. (e) If a single sample were to be selected for both purposes, the larger of the two sample sizes ($n = 121$) should be used.

8.60 (a) $0.2459 \leq \pi \leq 0.3741$. (b) $3.22 \leq \mu \leq \$3.78$.

(c) $\$17,581.68 \leq \mu \leq \$18,418.32$.

8.62 (a) $\$36.66 \leq \mu \leq \40.42 . (b) $0.2027 \leq \pi \leq 0.3973$.

(c) $n = 110$. (d) $n = 423$. (e) If a single sample were to be selected for both purposes, the larger of the two sample sizes ($n = 423$) should be used.

8.64 (a) $0.4643 \leq \pi \leq 0.6690$. (b) $\$136.28 \leq \mu \leq \502.21 .

8.66 (a) $8.41 \leq \mu \leq 8.43$. (b) With 95% confidence, the population mean width of troughs is somewhere between 8.41 and 8.43 inches. (c) The assumption is valid as the width of the troughs is approximately normally distributed.

8.68 (a) $0.2425 \leq \mu \leq 0.2856$. (b) $0.1975 \leq \mu \leq 0.2385$. (c) The amounts of granule loss for both brands are skewed to the right, but the sample sizes are large enough. (d) Because the two confidence intervals do not overlap, you can conclude that the mean granule loss of Boston shingles is higher than that of Vermont shingles.

CHAPTER 9

9.2 Because $Z_{STAT} = +2.21 > 1.96$, reject H_0 .

9.4 Reject H_0 if $Z_{STAT} < -2.58$ or if $Z_{STAT} > 2.58$.

9.6 p -value = 0.0456.

9.8 p -value = 0.1676.

9.10 H_0 : Defendant is guilty; H_1 : Defendant is innocent. A Type I error would be not convicting a guilty person. A Type II error would be convicting an innocent person.

9.12 H_0 : $\mu = 20$ minutes. 20 minutes is adequate travel time between classes. H_1 : $\mu \neq 20$ minutes. 20 minutes is not adequate travel time between classes.

$$\text{(a)} Z_{STAT} = \frac{7,250 - 7,500}{\sqrt{\frac{1,000}{64}}} = -2.0. \text{ Because } Z_{STAT} = -2.00$$

< -1.96 , reject H_0 . (b) p -value = 0.0456. (c) $7,005 \leq \mu \leq 7,495$. (d) The conclusions are the same.

9.16 (a) Because $-2.58 < Z_{STAT} = -1.7678 < 2.58$, do not reject H_0 . (b) p -value = 0.0771. (c) $0.9877 \leq \mu \leq 1.0023$. (d) The conclusions are the same.

9.18 $t_{STAT} = 2.00$.

9.20 ± 2.1315 .

9.22 No, you should not use a t test because the original population is left-skewed, and the sample size is not large enough for the t test to be valid.

9.24 (a) $t_{STAT} = (3.57 - 3.70)/0.8/\sqrt{64} = -1.30$. Because $-1.9983 < t_{STAT} = -1.30 < 1.9983$ and p -value = 0.1984 > 0.05, there is no evidence that the population mean waiting time is different from 3.7 minutes. (b) Because $n = 64$, the sampling distribution of the t test statistic is approximately normal. In general, the t test is appropriate for this sample size except for the case where the population is extremely skewed or bimodal.

9.26 (a) $-1.9842 < t_{STAT} = 1.4545 < 1.9842$. There is no evidence that the population mean savings for all showrooms is different from \$50.

(b) p -value = 0.1490 > 0.05. The probability of getting a t_{STAT} statistic greater than +1.4545 or less than -1.4545, given that the null hypothesis is true, is 0.1490.

9.28 (a) Because $-2.1448 < t_{STAT} = 1.6344 < 2.1448$, do not reject H_0 . There is not enough evidence to conclude that the mean amount spent for lunch at a fast food restaurant, is different from \$6.50. (b) The p -value is 0.1245. If the population mean is \$6.50, the probability of observing a sample of nine customers that will result in a sample mean farther away from the hypothesized value than this sample is 0.1245. (c) The distribution of the amount spent is normally distributed. (d) With a sample size of 15, it is difficult to evaluate the assumption of normality. However, the distribution may be fairly symmetric because the mean and the median are close in value. Also, the boxplot appears only slightly skewed so the normality assumption does not appear to be seriously violated.

9.30 (a) Because $-2.0096 < t_{STAT} = 0.114 < 2.0096$, do not reject H_0 . There is no evidence that the mean amount is different from 2 liters.

(b) p -value = 0.9095. (d) Yes, the data appear to have met the normality assumption. (e) The amount of fill is decreasing over time so the values are not independent. Therefore, the t test is invalid.

9.32 (a) Because $t_{STAT} = -5.9355 < -2.0106$, reject H_0 . There is enough evidence to conclude that mean widths of the troughs is different from 8.46 inches. **(b)** The population distribution is normal. **(c)** Although the distribution of the widths is left-skewed, the large sample size means that the validity of the t test is not seriously affected. The large sample size allows you to use the t distribution.

9.34 (a) Because $-2.68 < t_{STAT} = 0.094 < 2.68$, do not reject H_0 . There is no evidence that the mean amount is different from 5.5 grams. **(b)** $5.462 \leq \mu \leq 5.542$. **(c)** The conclusions are the same.

9.36 p -value = 0.0228.

9.38 p -value = 0.0838.

9.40 p -value = 0.9162.

9.42 $t_{STAT} = 2.7638$.

9.44 $t_{STAT} = -2.5280$.

9.46 (a) $t_{STAT} = 2.7273 > 1.6604$. There is evidence that the population mean bus miles is greater than 3,900 miles. **(b)** p -value = 0.0038 < 0.05. The probability of getting a t_{STAT} statistic greater than 2.7273 given that the null hypothesis is true, is 0.0038.

9.48 (a) $t_{STAT} = (23.05 - 25)/16.83/\sqrt{355} = -2.1831$. Because $t_{STAT} = -2.1831 > -2.3369$, do not reject H_0 . p -value = 0.0148 > 0.01, do not reject H_0 . **(b)** The probability of getting a sample mean of 23.05 minutes or less if the population mean is 25 minutes is 0.0148.

9.50 (a) $t_{STAT} = 4.1201 > 2.3974$. There is evidence that the population mean one-time gift donation is greater than \$70. **(b)** The probability of getting a sample mean of \$75 or more if the population mean is \$70 is 0.0001.

9.52 p = 0.22.

9.54 Do not reject H_0 .

9.56 (a) $Z_{STAT} = 1.1685$, p -value = 0.1213. Because $Z_{STAT} = 1.1685 < 1.645$ or $0.1213 > 0.05$, do not reject H_0 . There is no evidence to show that more than 20.3% of students at your university use the Mozilla Foundation web browser. **(b)** $Z_{STAT} = 2.3370$, p -value = 0.0097. Because $Z_{STAT} = 2.3370 > 1.645$, reject H_0 . There is evidence to show that more than 20.3% of students at your university use the Mozilla Foundation web browser. **(c)** The sample size had a major effect on being able to reject the null hypothesis. **(d)** You would be very unlikely to reject the null hypothesis with a sample of 20.

9.58 $H_0: \pi = 0.35$; $H_1: \pi \neq 0.35$. Decision rule: If $Z_{STAT} > 1.96$ or $Z_{STAT} < -1.96$, reject H_0 .

$$p = \frac{328}{801} = 0.4095$$

Test statistic:

$$Z_{STAT} = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} = \frac{0.4095 - 0.35}{\sqrt{\frac{0.4095(1 - 0.4095)}{801}}} = 3.5298.$$

Because $Z_{STAT} = 3.5298 > 1.96$ or p -value = 0.0004 < 0.05, reject H_0 and conclude that there is evidence that the proportion of all LinkedIn members who plan to spend at least \$1,000 on consumer electronics in the coming year is different from 35%.

9.60 (a) $H_0: \pi \geq 0.31$. The proportion who respond that shared organizational goals and objectives linking the team is the supported driver of alignment is greater than or equal to 0.31. $H_1: \pi < 0.31$. The proportion who respond that shared organizational goals and objectives linking the team is the supported driver of alignment is less than 0.31. **(b)** $Z_{STAT} = -0.6487 > -1.645$; p -value = 0.2583. Because $Z_{STAT} = -0.6487 > -1.645$ or p -value = 0.2583 > 0.05, reject H_0 . There is insufficient evidence that the proportion who respond that shared organizational goals and objectives linking the team is the supported driver of alignment is less than 0.31.

9.70 (a) Concluding that a firm will go bankrupt when it will not. **(b)** Concluding that a firm will not go bankrupt when it will go bankrupt. **(c)** Type I. **(d)** If the revised model results in more moderate or large Z scores, the probability of committing a Type I error will increase. Many more of the firms will be predicted to go bankrupt than will go bankrupt. On the other hand, the revised model that results in more moderate or large Z scores will lower the probability of committing a Type II error because few firms will be predicted to go bankrupt than will actually go bankrupt.

9.72 (a) Because $t_{STAT} = 3.3197 > 2.0010$, reject H_0 . **(b)** p -value = 0.0015. **(c)** Because $Z_{STAT} = 0.2582 < 1.645$, do not reject H_0 . **(d)** Because $-2.0010 < t_{STAT} = -1.1066 < 2.0010$, do not reject H_0 . **(e)** Because $Z_{STAT} = 2.3238 > 1.645$, reject H_0 .

9.74 (a) Because $t_{STAT} = -1.69 > -1.7613$, do not reject H_0 . **(b)** The data are from a population that is normally distributed. **(d)** With the exception of one extreme value, the data are approximately normally distributed. **(e)** There is insufficient evidence to state that the waiting time is less than five minutes.

9.76 (a) Because $t_{STAT} = -1.47 > -1.6896$, do not reject H_0 . **(b)** p -value = 0.0748. If the null hypothesis is true, the probability of obtaining a t_{STAT} of -1.47 or more extreme is 0.0748. **(c)** Because $t_{STAT} = -3.10 < -1.6973$, reject H_0 . **(d)** p -value = 0.0021. If the null hypothesis is true, the probability of obtaining a t_{STAT} of -3.10 or more extreme is 0.0021. **(e)** The data in the population are assumed to be normally distributed. **(g)** Both boxplots suggest that the data are skewed slightly to the right, more so for the Boston shingles. However, the very large sample sizes mean that the results of the t test are relatively insensitive to the departure from normality.

9.78 (a) $t_{STAT} = -3.2912$, reject H_0 . **(b)** p -value = 0.0012. **(c)** $t_{STAT} = -7.9075$, reject H_0 . **(d)** p -value = 0.0000. **(e)** Because of the large sample sizes, you do not need to be concerned with the normality assumption.

CHAPTER 10

10.2 (a) $t = 3.8959$. **(b)** $df = 21$. **(c)** 2.5177. **(d)** Because $t_{STAT} = t_{STAT} = 3.8959 > 2.5177$, reject H_0 .

10.4 $3.73 \leq \mu_1 - \mu_2 \leq 12.27$.

10.6 Because $t_{STAT} = 2.6762 < 2.9979$ or p -value = 0.0158 > 0.01, do not reject H_0 . There is no evidence of a difference in the means of the two populations.

10.8 (a) Because $t_{STAT} = 2.8990 > 1.6620$ or p -value = 0.0024 < 0.05, reject H_0 . There is evidence that the mean amount of Walker Crisps eaten by children who watched a commercial featuring a long-standing sports celebrity endorser is higher than for those who watched a commercial for an alternative food snack. **(b)** $3.4616 \leq \mu_1 - \mu_2 \leq 18.5384$. **(c)** The

results cannot be compared because (a) is a one-tail test and (b) is a confidence interval that is comparable only to the results of a two-tail test.

- 10.10 (a)** $H_0: \mu_1 = \mu_2$, where Populations: 1 = Southeast, 2 = Gulf Coast. $H_1: \mu_1 \neq \mu_2$. Decision rule: $df = 29$. If $t_{STAT} < -2.0452$ or $t_{STAT} > 2.0452$, reject H_0 .

Test statistic:

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)(S_1^2) + (n_2 - 1)(S_2^2)}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{(13)(40.9906^2) + (16)(32.0906^2)}{14 + 17} = 1,321.3748 \\ t_{STAT} &= \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \\ &= \frac{(40.0714 - 27.9412) - 0}{\sqrt{1,321.3748 \left(\frac{1}{14} + \frac{1}{17} \right)}} = 0.9246. \end{aligned}$$

Decision: Because $-2.0452 < t_{STAT} = 0.9246 < 2.0452$, do not reject H_0 . There is not enough evidence to conclude that the mean number of partners between the Southeast and Gulf Coast is different. **(b)** p -value = 0.3628. **(c)** In order to use the pooled-variance t test, you need to assume that the populations are normally distributed with equal variances.

- 10.12 (a)** Because $t_{STAT} = -4.1343 < -2.0484$, reject H_0 . **(b)** p -value = 0.0003. **(c)** The populations of waiting times are approximately normally distributed. **(d)** $-4.2292 \leq \mu_1 - \mu_2 \leq -1.4268$.
- 10.14 (a)** Because $t_{STAT} = -1.4707 > -2.0484$, do not reject H_0 . There is insufficient evidence of a difference in the mean time to start a business between developed and emerging countries. **(b)** p -value = 0.1525. The probability that two samples have a mean difference of 11.20 or more is 0.1525 if there is no difference in the mean time to start a business between developed and emerging countries. **(c)** You need to assume that the population distribution of the time to start a business of both developed and emerging countries is normally distributed. **(d)** $-26.7996 \leq \mu_1 - \mu_2 \leq 4.3996$.

- 10.16 (a)** Because $t_{STAT} = -2.0036 < -2.0017$ or p -value = 0.0498 < 0.05, reject H_0 . There is evidence of a difference in the mean Facebook time per day between males and females. **(b)** You must assume that each of the two independent populations is normally distributed.

10.18 $df = 19$.

- 10.20 (a)** $t_{STAT} = (-1.5566)/(1.424)/\sqrt{9} = -3.2772$. Because $t_{STAT} = -3.2772 < -2.306$ or p -value = 0.0112 < 0.05, reject H_0 . There is enough evidence of a difference in the mean summated ratings between the two brands. **(b)** You must assume that the distribution of the differences between the two ratings is approximately normal. **(c)** p -value = 0.0112. The probability of obtaining a mean difference in ratings that results in a test statistic that deviates from 0 by 3.2772 or more in either direction is 0.0112 if there is no difference in the mean summated ratings between the two brands. **(d)** $-2.6501 \leq \mu_D \leq -0.4610$. You are 95% confident that the mean difference in summated ratings between brand A and brand B is somewhere between -2.6501 and -0.4610 .

- 10.22 (a)** Because $t_{STAT} = 1.7948 > 1.6939$ reject H_0 . There is evidence to conclude that the mean at Super Target is higher than at Walmart.

- (b)** You must assume that the distribution of the differences between the prices is approximately normal. **(c)** p -value = 0.0411. The likelihood that you will obtain a t_{STAT} statistic greater than 1.7948 if the mean price at Super Target is not greater than Walmart is 0.0411.

- 10.24 (a)** Because $t_{STAT} = 1.8425 < 1.943$, do not reject H_0 . There is not enough evidence to conclude that the mean bone marrow microvessel density is higher before the stem cell transplant than after the stem cell transplant. **(b)** p -value = 0.0575. The probability that the t statistic for the mean difference in microvessel density is 1.8425 or more is 5.75% if the mean density is not higher before the stem cell transplant than after the stem cell transplant. **(c)** $-28.26 \leq \mu_D \leq 200.55$. You are 95% confident that the mean difference in bone marrow microvessel density before and after the stem cell transplant is somewhere between -28.26 and 200.55 . **(d)** That the distribution of the difference before and after the stem cell transplant is normally distributed.

- 10.26 (a)** Because $t_{STAT} = -9.3721 < -2.4258$, reject H_0 . There is evidence that the mean strength is lower at two days than at seven days. **(b)** The population of differences in strength is approximately normally distributed. **(c)** $p = 0.000$.

- 10.28 (a)** Because $-2.58 \leq Z_{STAT} = -0.58 \leq 2.58$, do not reject H_0 . **(b)** $-0.273 \leq \pi_1 - \pi_2 \leq 0.173$.

- 10.30 (a)** $H_0: \pi_1 \leq \pi_2$. $H_1: \pi_1 > \pi_2$. Populations: 1 = social media recommendation, 2 = web browsing. **(b)** Because $Z_{STAT} = 1.5507 < 1.6449$ or p -value = 0.0605 > 0.05, do not reject H_0 . There is insufficient evidence to conclude that the population proportion of those who recalled the brand is greater for those who had a social media recommendation than for those who did web browsing. **(c)** No, the result in (b) makes it inappropriate to claim that the population proportion of those who recalled the brand is greater for those who had a social media recommendation than for those who did web browsing.

- 10.32 (a)** $H_0: \pi_1 = \pi_2$. $H_1: \pi_1 \neq \pi_2$. Decision rule: If $|Z_{STAT}| > 2.58$, reject H_0 .

Test statistic: $\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{930 + 230}{1,000 + 1,000} = 0.58$

$$Z_{STAT} = \frac{(p_1 - p_2) - (\pi_2 - \pi_2)}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0.93 - 0.23) - 0}{\sqrt{0.58(1 - 0.58)\left(\frac{1}{1,000} + \frac{1}{1,000}\right)}} = 31.7135$$

- $Z_{STAT} = 31.7135 > 2.58$, reject H_0 . There is evidence of a difference in the proportion of Superbanked and Unbanked with respect to the proportion that use credit cards. **(b)** p -value = 0.0001. The probability of obtaining a difference in proportions that gives rise to a test statistic below -31.7135 or above $+31.7135$ is 0.0000 if there is no difference in the proportion of Superbanked and Unbanked who use credit cards. **(c)** $0.6599 \leq (\pi_1 - \pi_2) \leq 0.7401$. You are 99% confident that the difference in the proportion of Superbanked and Unbanked who use credit cards is between 0.6599 and 0.7401.

- 10.34 (a)** Because $Z_{STAT} = 3.8512 > 1.96$, reject H_0 . There is evidence of a difference in the proportion of gamification-user sales organizations and non-gamification-user sales organizations that provide mobile access to CRM. **(b)** p -value = 0.0001. The probability of obtaining a difference in proportions that is 0.336 or more in either direction is 0.0001 if there is no difference between the proportion of gamification-user sales organizations and non-gamification-user sales organizations that provide mobile access to CRM.

- 10.36 (a)** 2.20. **(b)** 2.57. **(c)** 3.50.

- 10.38 (a)** Population B: $S^2 = 25$. **(b)** 1.5625.

10.40 $df_{\text{numerator}} = 24$, $df_{\text{denominator}} = 24$.

10.42 Because $F_{\text{STAT}} = 1.2109 < 2.27$, do not reject H_0 .

10.44 (a) Because $F_{\text{STAT}} = 1.2995 < 3.18$, do not reject H_0 . **(b)** Because $F_{\text{STAT}} = 1.2995 < 2.62$, do not reject H_0 .

10.46 (a) $H_0: \sigma_1^2 = \sigma_2^2$. $H_1: \sigma_1^2 \neq \sigma_2^2$.

Decision rule: If $F_{\text{STAT}} > 2.8506$, reject H_0 .

$$\text{Test statistic: } F_{\text{STAT}} = \frac{S_1^2}{S_2^2} = \frac{(40.9906)^2}{(32.0906)^2} = 1.6316.$$

Decision: Because $F_{\text{STAT}} = 1.6316 < 2.8506$, do not reject H_0 . There is insufficient evidence to conclude that the two population variances are different. **(b)** $p\text{-value} = 0.3509$. **(c)** The test assumes that each of the two populations is normally distributed. **(d)** Based on (a) and (b), a pooled-variance t test should be used.

10.48 (a) Because $F_{\text{STAT}} = 3.8179 < 4.0721$ or $p\text{-value} = 0.0609 < 0.05$, do not reject H_0 . There is no evidence of a difference in the variability of the battery life between the two types of digital cameras. **(b)** $p\text{-value} = 0.0609$. The probability of obtaining a sample that yields a test statistic more extreme than 3.8179 is 0.0609 if there is no difference in the two population variances. **(c)** The test assumes that each of the two populations are normally distributed. The boxplots appear fairly symmetrical and the skewness and kurtosis statistics are not dramatically different from 0. Thus, the distributions do not appear to be substantially different from a normal distribution. **(d)** Based on (a) and (b), a pooled-variance t test should be used.

10.50 Because $F_{\text{STAT}} = 1.1583 < 4.8232$, or $p\text{-value} = 0.8658 > 0.05$, do not reject H_0 . There is insufficient evidence of a difference in the variance of the yield in the two cities.

10.58 (a) Because $F_{\text{STAT}} = 1.2221 < 1.7462$, or $p\text{-value} = 0.4688 > 0.05$, do not reject H_0 . There is not enough evidence of a difference in the variance of the salary of Black Belts and Green Belts. **(b)** The pooled-variance t test. **(c)** Because $t_{\text{STAT}} = 4.2412 > 1.6554$ or $p\text{-value} = 0.0000 < 0.05$, reject H_0 . There is evidence that the mean salary of Black Belts is greater than the mean salary of Green Belts.

10.60 (a) Because $F_{\text{STAT}} = 1.5625 < F_\alpha = 1.6854$, do not reject H_0 . There is not enough evidence to conclude that there is a difference between the variances in the talking time per month between women and men. **(b)** It is more appropriate to use a pooled-variance t test. Using the pooled-variance t test, because $t_{\text{STAT}} = 11.1196 > 2.6009$, reject H_0 . There is enough evidence of a difference in the mean talking time per month between women and men. **(c)** Because $F_{\text{STAT}} = 1.44 < 1.6854$, do not reject H_0 . There is not enough evidence to conclude that there is a difference between the variances in the number of text messages sent per month between women and men. **(d)** Using the pooled-variance t test, because $t_{\text{STAT}} = 8.2456 > 2.6009$, reject H_0 . There is enough evidence of a difference in the mean number of text messages sent per month between women and men.

10.62 (a) Because $t_{\text{STAT}} = 3.3282 > 1.8595$, reject H_0 . There is enough evidence to conclude that the introductory computer students required more than a mean of 10 minutes to write and run a program in VB.NET. **(b)** Because $t_{\text{STAT}} = 1.3636 < 1.8595$, do not reject H_0 . There is not enough evidence to conclude that the introductory computer students required more than a mean of 10 minutes to write and run a program in VB.NET. **(c)** Although the mean time necessary to complete the assignment increased from 12 to 16 minutes as a result of the increase in one

data value, the standard deviation went from 1.8 to 13.2, which reduced the value of t statistic. **(d)** Because $F_{\text{STAT}} = 1.2308 < 3.8549$, do not reject H_0 . There is not enough evidence to conclude that the population variances are different for the Introduction to Computers students and computer majors. Hence, the pooled-variance t test is a valid test to determine whether computer majors can write a VB.NET program in less time than introductory students, assuming that the distributions of the time needed to write a VB.NET program for both the Introduction to Computers students and the computer majors are approximately normally distributed. Because $t_{\text{STAT}} = 4.0666 > 1.7341$, reject H_0 . There is enough evidence that the mean time is higher for Introduction to Computers students than for computer majors. **(e)** $p\text{-value} = 0.000362$. If the true population mean amount of time needed for Introduction to Computer students to write a VB.NET program is no more than 10 minutes, the probability of observing a sample mean greater than the 12 minutes in the current sample is 0.0362%. Hence, at a 5% level of significance, you can conclude that the population mean amount of time needed for Introduction to Computer students to write a VB.NET program is more than 10 minutes. As illustrated in (d), in which there is not enough evidence to conclude that the population variances are different for the Introduction to Computers students and computer majors, the pooled-variance t test performed is a valid test to determine whether computer majors can write a VB.NET program in less time than introductory students, assuming that the distribution of the time needed to write a VB.NET program for both the Introduction to Computers students and the computer majors are approximately normally distributed.

10.64 From the boxplot and the summary statistics, both distributions are approximately normally distributed. $F_{\text{STAT}} = 1.056 < 1.89$.

There is insufficient evidence to conclude that the two population variances are significantly different at the 5% level of significance. $t_{\text{STAT}} = -5.084 < -1.99$. At the 5% level of significance, there is sufficient evidence to reject the null hypothesis of no difference in the mean life of the bulbs between the two manufacturers. You can conclude that there is a significant difference in the mean life of the bulbs between the two manufacturers.

10.66 (a) Because $Z_{\text{STAT}} = -3.6911 < -1.96$, reject H_0 . There is enough evidence to conclude that there is a difference in the proportion of men and women who order dessert. **(b)** Because $Z_{\text{STAT}} = 6.0873 > 1.96$, reject H_0 . There is enough evidence to conclude that there is a difference in the proportion of people who order dessert based on whether they ordered a beef entree.

10.68 The normal probability plots suggest that the two populations are not normally distributed. An F test is inappropriate for testing the difference in the two variances. The sample variances for Boston and Vermont shingles are 0.0203 and 0.015, respectively. Because $t_{\text{STAT}} = 3.015 > 1.967$ or $p\text{-value} = 0.0028 < \alpha = 0.05$, reject H_0 . There is sufficient evidence to conclude that there is a difference in the mean granule loss of Boston and Vermont shingles.

CHAPTER 11

11.2 (a) $SSW = 150$. **(b)** $MSA = 15$. **(c)** $MSW = 5$. **(d)** $F_{\text{STAT}} = 3$.

11.4 (a) 2. **(b)** 18. **(c)** 20.

11.6 (a) Reject H_0 if $F_{\text{STAT}} > 2.95$; otherwise, do not reject H_0 .

(b) Because $F_{\text{STAT}} = 4 > 2.95$, reject H_0 . **(c)** The table does not have 28 degrees of freedom in the denominator, so use the next larger critical value, $Q_\alpha = 3.90$. **(d)** Critical range = 6.166.

11.8 (a) $H_0: \mu_A = \mu_B = \mu_C = \mu_D$ and $H_1:$ At least one mean is different.

$$MSA = \frac{SSA}{c - 1} = \frac{7,854,648}{3} = 2,618,216.2.$$

$$MSW = \frac{SSW}{n - c} = \frac{15,137,801.8}{36} = 420,494.4944.$$

$$F_{STAT} = \frac{MSA}{MSW} = \frac{2,618,216.2}{420,494.4944} = 6.2265.$$

$$F_{0.05,3,36} = 2.8663.$$

Because the p -value is 0.0016 and $F_{STAT} = 6.2265 > 2.8663$, reject H_0 . There is sufficient evidence of a difference in the mean import cost across the four global regions. **(b)** Critical range = $Q_\alpha \sqrt{\frac{MSW}{2} \left(\frac{1}{n_j} + \frac{1}{n_{j'}} \right)}$

$$= 3.79 \sqrt{\frac{420,494.4944}{2} \left(\frac{1}{10} + \frac{1}{10} \right)} = 205.0596.$$

From the Tukey-Kramer procedure, there is a difference in the mean import cost between the East Asia and Pacific region and each of the other regions. None of the other regions are different. **(c)** ANOVA output for Levene's test for homogeneity of variance:

$$MSA = \frac{SSA}{c - 1} = \frac{1,387,853.3}{3} = 462,617.7667$$

$$MSW = \frac{SSW}{n - c} = \frac{7,654,784.8}{36} = 212,632.9111$$

$$F_{STAT} = \frac{MSA}{MSW} = \frac{462,617.7667}{212,632.9111} = 2.1757$$

$$F_{0.05,3,36} = 2.8663$$

Because p -value = 0.1078 > 0.05 and $F_{STAT} = 2.1757 < 2.8663$, do not reject H_0 . There is insufficient evidence to conclude that the variances in the import cost are different. **(d)** From the results in (a) and (b), the mean import cost for the East Asia and Pacific region is lower than for the other regions.

11.10 (a) Because $F_{STAT} = 12.56 > 2.76$, reject H_0 . **(b)** Critical range = 4.67. Advertisements A and B are different from Advertisements C and D. Advertisement E is only different from Advertisement D.

(c) Because $F_{STAT} = 1.927 < 2.76$, do not reject H_0 . There is no evidence of a significant difference in the variation in the ratings among the five advertisements. **(d)** The advertisements underselling the pen's characteristics had the highest mean ratings, and the advertisements overselling the pen's characteristics had the lowest mean ratings. Therefore, use an advertisement that undersells the pen's characteristics and avoid advertisements that oversell the pen's characteristics.

11.12 (a)

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
Among groups	2	1.879	0.9395	8.7558
Within groups	297	31.865	0.1073	
Total	299	33.744		

(b) Since $F_{STAT} = 8.7558 > 3.00$, reject H_0 . There is evidence of a difference in the mean soft-skill score of the different groups.

(c) Group 1 versus group 2: $0.072 <$ Critical range = 0.1092; group 1 versus group 3: $0.181 > 0.1056$; group 2 versus group 3: $0.109 < 0.1108$. There is evidence of a difference in the mean soft-skill score between those who had no coursework in leadership and those who had a degree in leadership.

11.14 (a) Because $F_{STAT} = 53.03 > 2.92$, reject H_0 . **(b)** Critical range = 5.27 (using 30 degrees of freedom). Designs 3 and 4 are different from designs 1 and 2. Designs 1 and 2 are different from each other. **(c)** The assumptions are that the samples are randomly and independently selected (or randomly assigned), the original populations of distances are approximately normally distributed, and the variances are equal. **(d)** Because $F_{STAT} = 2.093 < 2.92$, do not reject H_0 . There is insufficient evidence of a difference in the variation in the distance among the four designs. **(e)** The manager should choose design 3 or 4.

11.16 (a) $SSE = 75$. **(b)** $MSA = 15$, $MSBL = 12.5$, $MSE = 3.125$. **(c)** $F_{STAT} = 4.8$. **(d)** $F_{STAT} = 4$.

11.18 (a) df numerator = 5, df denominator = 24. **(b)** $Q_\alpha = 4.17$. **(c)** Critical range = 2.786.

11.20 (a) $MSE = 3$, $SSE = 36$, because $F_{STAT} = 4 < F_{0.01,6,12} = 4.82$, do not reject H_0 .

11.22 (a) Because $F_{STAT} = 5.185 > 3.07$, reject H_0 . **(b)** Because $F_{STAT} = 5 > 2.49$, reject H_0 .

11.24 (a)

Anova: Two-Factor Without Replication						
SUMMARY	Count	Sum	Average	Variance		
Verizon FIOS	3	228	76	4		
WOW	3	235	78.33333	20.33333		
Bright House Networks	3	219	73	13		
A T & T U-verse	3	208	69.33333	5.333333		
Cox	3	211	70.33333	12.33333		
SuddenLink	3	215	71.66667	66.33333		
Cablevision/Optimum	3	206	68.66667	32.33333		
Insight	3	204	68	21		
RCN	3	198	66	21		
Comcast	3	195	65	21		
TimeWarner	3	196	65.33333	30.33333		
Charter	3	193	64.33333	37.33333		
Mediacom	3	174	58	21		
TV	13	843	64.84615	34.14103		
Phone	13	959	73.76923	27.02564		
Internet	13	880	67.69231	30.39744		
ANOVA						
Source of Variation	ss	df	MS	F	P-value	F crit
Rows	1,028.256	12	85.68803	29.16509	0.0000	2.18338
Columns	540.1538	2	270.0769	91.92436	0.0000	3.402826
Error	70.51282	24	2.938034			
Total	1,638.923	38				

$F_{STAT} = 91.9243 > 3.4028$, p -value is 0.0000, reject H_0 . There is evidence of a difference in the mean score between TV, phone, and Internet.

(b) Critical range = $3.53 \sqrt{\frac{2.938034}{13}} = 1.6781$; TV – phone: $64.8462 - 73.7692 = -8.9230$; TV – Internet: $64.8462 - 67.6923 = -2.8461$; Phone – Internet: $73.7692 - 67.6923 = 6.0769$. The mean for TV is significantly lower than for phone and for Internet. Then mean for phone is significantly higher than for Internet.

11.26 (a) $H_0: \mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{.4}$ $H_1:$ Not all μ_j are equal where $j = 1, 2, 3, 4$ $F_{STAT} = 99.8046 > 2.8115$, p -value = 0.0000 < 0.05, reject H_0 . There is evidence of a difference in the mean rates for the different investments. **(b)** The assumptions needed are (i) samples are randomly and independently drawn, (ii) populations are normally distributed,

(iii) populations have equal variances, and (iv) no interaction effect between treatments and blocks. **(c)**, **(d)** $H_0: \mu_1 = \mu_2 = \dots = \mu_{16}, H_0:$ Not all μ_i are equal where $i = 1, 2, \dots, 16$, $F_{STAT} = 12.2177 > 1.8949$, and $p\text{-value} = 0.0000$, reject H_0 . There is evidence of a significant block effect in this experiment. The blocking has been advantageous in reducing the experimental error.

11.28 (a) Because $F_{STAT} = 268.26 > 3.114$, reject H_0 . There is enough evidence to conclude that there is a difference in the mean compressive strength after 2, 7, and 28 days. **(b)** Critical range = 0.1651. At the 0.05 level of significance, all of the comparisons are significant. **(c)** $RE = 2.558$. **(e)** The compressive strength of the concrete increases over the three time periods.

11.30 (a) 40. **(b)** 60 and 55. **(c)** 10. **(d)** 10.

11.32 (a) Because $F_{STAT} = 6.00 > 3.35$, reject H_0 . **(b)** Because $F_{STAT} = 5.50 > 3.35$, reject H_0 . **(c)** Because $F_{STAT} = 1.00 < 2.73$, do not reject H_0 .

11.34 $df_B = 4$, $df_{TOTAL} = 44$, $SSA = 160$, $SSAB = 80$, $SSE = 150$, $SST = 610$, $MSB = 55$, $MSE = 5$. For A: $F_{STAT} = 16$. For B: $F_{STAT} = 11$. For AB: $F_{STAT} = 2$.

11.36 (a) Because $F_{STAT} = 3.4032 < 4.3512$, do not reject H_0 . **(b)** Because $F_{STAT} = 1.8496 < 4.3512$, do not reject H_0 . **(c)** Because $F_{STAT} = 9.4549 > 4.3512$ reject H_0 . **(e)** Die diameter has a significant effect on density, but die temperature does not. However, the cell means plot shows that the density seems higher with a 3 mm die diameter at 155 C but that there is little difference in density with a 4 mm die diameter. This interaction is not significant at the 0.05 level of significance.

11.38 (a) H_0 : There is no interaction between brand and water temperature. H_1 : There is an interaction between brand and water temperature.

Because $F_{STAT} = \frac{253.1552}{12.2199} = 20.7167 > 3.555$ or the $p\text{-value} = 0.0000214 < 0.05$, reject H_0 . There is evidence of interaction between brand of pain reliever and temperature of the water. **(b)** Because there is an interaction between brand and the temperature of the water, it is inappropriate to analyze the main effect due to brand. **(c)** Because there is an interaction between brand and the temperature of the water, it is inappropriate to analyze the main effect due to water temperature. **(e)** The difference in the mean time a tablet took to dissolve in cold and hot water depends on the brand, with Alka-Seltzer having the largest difference and Equate the smallest difference.

11.40 (a) $F_{STAT} = 0.1523$, $p\text{-value} = 0.9614 > 0.05$, do not reject H_0 . There is not enough evidence to conclude that there is an interaction between the brake discs and the gauges. **(b)** $F_{STAT} = 7.7701$, $p\text{-value}$ is virtually 0 < 0.05, reject H_0 . There is sufficient evidence to conclude that there is an effect due to brake discs. **(c)** $F_{STAT} = 0.1465$, $p\text{-value} = 0.7031 > 0.05$, do not reject H_0 . There is inadequate evidence to conclude that there is an effect due to the gauges. **(d)** From the plot, there is no obvious interaction between brake discs and gauges. **(e)** There is no obvious difference in mean temperature across the gauges. It appears that Part 1 has the lowest, Part 3 the second lowest, and Part 2 has the highest average temperature.

11.52 (a) Because $F_{STAT} = 0.0111 < 2.9011$, do not reject H_0 . **(b)** Because $F_{STAT} = 0.8096 < 4.1491$, do not reject H_0 . **(c)** Because $F_{STAT} = 5.1999 > 2.9011$, reject H_0 . **(e)** Critical range = 3.56. Only the means of Suppliers 1 and 2 are different. You can conclude that the mean tensile strength is lower for Supplier 1 than for Supplier 2, but there are no statistically significant differences between Suppliers 1 and 3, Suppliers 1 and 4, Suppliers 2 and 3, Suppliers 2 and 4, and Suppliers 3 and 4. **(f)** $F_{STAT} = 5.6998 > 2.8663$ ($p\text{-value} = 0.0027 < 0.05$).

There is evidence that the mean strength of suppliers is different. Critical range = 3.359. Supplier 1 has a mean strength that is less than suppliers 2 and 3.

11.54 (a) Because $F_{STAT} = 0.075 < 3.68$, do not reject H_0 . **(b)** Because $F_{STAT} = 4.09 > 3.68$, reject H_0 . **(c)** Critical range = 1.489. Breaking strength is significantly different between 30 and 50 psi.

11.56 (a) Because $F_{STAT} = 0.1899 < 4.1132$, do not reject H_0 . There is insufficient evidence to conclude that there is any interaction between type of breakfast and desired time. **(b)** Because $F_{STAT} = 30.4434 > 4.1132$, reject H_0 . There is sufficient evidence to conclude that there is an effect due to type of breakfast. **(c)** Because $F_{STAT} = 12.4441 > 4.1132$, reject H_0 . There is sufficient evidence to conclude that there is an effect due to desired time. **(e)** At the 5% level of significance, both the type of breakfast ordered and the desired time have an effect on delivery time difference. There is no interaction between the type of breakfast ordered and the desired time.

11.58 Interaction: $F_{STAT} = 0.2169 < 3.9668$ or $p\text{-value} = 0.6428 > 0.05$. There is insufficient evidence of an interaction between piece size and fill height. Piece size: $F_{STAT} = 842.2242 > 3.9668$ or $p\text{-value} = 0.0000 < 0.05$. There is evidence of an effect due to piece size. The fine piece size has a lower difference in coded weight. Fill height: $F_{STAT} = 217.0816 > 3.9668$ or $p\text{-value} = 0.0000 < 0.05$. There is evidence of an effect due to fill height. The low fill height has a lower difference in coded weight.

11.60 Population 1 = short term 2 = long term, 3 = world; One-year return: Levene test: Since the $p\text{-value} 0.5220 > 0.05$ do not reject H_0 . There is insufficient evidence to show a difference in the variance of the return among the three different types of bond funds at a 5% level of significance. Since the $p\text{-value}$ is 0.0000, reject H_0 . There is sufficient evidence to show a difference in the mean one-year returns among the three different types of bond funds at a 5% level of significance. Critical range = 2.8194. At the 5% level of significance, there is sufficient evidence that the mean one-year returns of the short-term bond funds are significantly lower than the others. Three-year return: Levene test: $F_{STAT} = 1.0557$. Since the $p\text{-value} = 0.3619 > 0.05$, do not reject H_0 . There is insufficient evidence to show a difference in the variance of return among the three different types of bond funds at a 5% level of significance. $F_{STAT} = 37.1365$. Since the $p\text{-value}$ is 0.0000, reject H_0 . There is sufficient evidence to show a difference in the mean three-year returns among the three different types of bond funds at a 5% level of significance. Critical range = 1.9394. At the 5% level of significance, there is sufficient evidence that the mean three-year returns of the short-term bond funds is significantly higher than the others. Also, the mean three-year returns of the long-term bond funds are significantly higher than that of the world bond funds.

CHAPTER 12

12.2 (a) For $df = 1$ and $\alpha = 0.05$, $\chi^2_\alpha = 3.841$. **(b)** For $df = 1$ and $\alpha = 0.025$, $\chi^2 = 5.024$. **(c)** For $df = 1$ and $\alpha = 0.01$, $\chi^2_\alpha = 6.635$.

12.4 (a) All $f_e = 25$. **(b)** Because $\chi^2_{STAT} = 4.00 > 3.841$, reject H_0 .

12.6 (a) $H_0: \pi_1 = \pi_2$. $H_1: \pi_1 \neq \pi_2$. **(b)** Because $\chi^2_{STAT} = 2.4045 < 3.841$, do not reject H_0 . There is insufficient evidence to conclude that the population proportion of those who recalled the brand is different for those who had a social media recommendation than for those who did web browsing. $p\text{-value} = 0.1210$. The probability of obtaining a test statistic of 2.4045 or larger when the null hypothesis is true is 0.1210. **(c)** You should not compare the results in (a) to those of Problem 10.30 (b) because that was a one-tail test.

12.8 (a) $H_0: \pi_1 = \pi_2$. $H_1: \pi_1 \neq \pi_2$. Because $\chi^2_{STAT} = (930 - 580)^2 / 580 + (70 - 420)^2 / 420 + (230 - 580)^2 / 580 + (770 - 420)^2 / 1,005.7471 > 6.635$, reject H_0 . There is evidence of a difference in the proportion of Superbanked and Unbanked with respect to the proportion that use credit cards. **(b)** $p\text{-value} = 0.0000$. The probability of obtaining a difference in proportions that gives rise to a test statistic above 1,005.7471 is 0.0000 if there is no difference in the proportion of Superbanked and Unbanked who use credit cards. **(c)** The results of (a) and (b) are exactly the same as those of Problem 10.32. The χ^2 in (a) and the Z in Problem 10.32 (a) satisfy the relationship that $\chi^2 = 1,005.7471 = Z^2 = (31.7135)^2$, and the $p\text{-value}$ in (b) is exactly the same as the $p\text{-value}$ computed in Problem 10.32 (b).

12.10 (b) Since $\chi^2_{STAT} = 14.8319 > 3.841$, reject H_0 . There is evidence that there is a significant difference between the gamification-user sales organizations and non-gamification-user sales organizations in the proportion that provide mobile access to CRM. **(c)** $p\text{-value} 0.0001$. The probability of obtaining a test statistic of 14.8319 or larger when the null hypothesis is true is 0.0001. **(d)** The results are identical since $(3.8512)^2 = 14.8319$.

12.12 (a) The expected frequencies for the first row are 20, 30, and 40. The expected frequencies for the second row are 30, 45, and 60.

(b) Because $\chi^2_{STAT} = 12.5 > 5.991$, reject H_0 .

12.14 (a) Because the calculated test statistic $\chi^2_{STAT} = 44.6503 > 11.0705$, reject H_0 and conclude that there is a difference in the proportion who buy lunch between the age groups. **(b)** The $p\text{-value}$ is virtually 0. The probability of a test statistic greater than 44.6503 or more is approximately 0 if there is no difference between the age groups in the proportion who buy lunch. **(c)** The 18–24 and 25–34 groups are different from the 45–54, 55–64, and 65+ groups, and the 35–44 group is different from the 65+ group.

12.16 (a) $H_0: \pi_1 = \pi_2 = \pi_3$. H_1 : At least one proportion differs.

f_0	f_e	$(f_0 - f_e)$	$(f_0 - f_e)^2 / f_e$
118	72	46	29.3889
82	128	-46	16.5313
72	72	0	0
128	128	0	0
26	72	-46	29.3889
174	128	46	16.5313
			91.8403

Decision rule: $df = (c - 1) = (3 - 1) = 2$. If $\chi^2_{STAT} > 5.9915$, reject H_0 .

Test statistic: $\chi^2_{STAT} = \sum_{\text{all cells}} \frac{(f_0 - f_e)^2}{f_e} = 91.8403$.

Decision: Because $\chi^2_{STAT} = 91.8403 > 5.9915$, reject H_0 . There is a significant difference in the age groups with respect to using a cellphone to access social networking. **(b)** $p\text{-value} = 0.0000$. The probability that the test statistic is greater than or equal to 91.8403 is 0.0000, if the null hypothesis is true.

(c)

Pairwise Comparisons	Critical Range	$ p_j - p_j' $
1 to 2	0.1189	0.23
1 to 3	0.1031	0.46
2 to 3	0.1014	0.23

There is a significant difference between all the groups. **(d)** Marketers can use this information to target their marketing to the 18- to 34-year-old group since they are more likely to use their cellphones to access social media.

12.18 (a) Because $\chi^2_{STAT} = 9.0485 > 5.9915$, reject H_0 . There is evidence of a difference in the percentage who use their cellphones while watching TV between the groups. **(b)** $p\text{-value} = 0.0108$.

(c) Group 1 versus group 2: $0.0215 < 0.0788$. Not significant. Group 1 versus group 3: $0.0905 > 0.0859$ significant. Group 2 versus group 3: $0.0691 > 0.06457$ significant. The rural group is different from the urban and suburban groups.

12.20 $df = (r - 1)(c - 1) = (3 - 1)(4 - 1) = 6$.

12.22 $\chi^2_{STAT} = 92.1028 > 16.919$, reject H_0 and conclude that there is evidence of a relationship between the type of dessert ordered and the type of entrée ordered.

12.24 H_0 : There is no relationship between the frequency of posting on Facebook and age. H_1 : There is a relationship between the frequency of posting on Facebook and age.

f_0	f_e	$(f_0 - f_e)$	$(f_0 - f_e)^2 / f_e$
20	8.895459	11.10454	13.8622
28	13.83738	14.16262	14.4955
32	19.43823	12.56177	8.1179
32	26.35692	5.643083	1.2082
22	27.83949	-5.83949	1.2249
16	34.59345	-18.5935	9.9937
6	25.03907	-19.0391	14.4768
22	17.90496	4.095037	0.936574
37	27.85216	9.147835	3.004538
46	39.12566	6.87434	1.207815
69	53.05174	15.94826	4.794318
66	56.0359	9.964097	1.771779
59	69.63041	-10.6304	1.622935
15	50.39916	-35.3992	24.86352
9	13.34319	-4.34319	1.413702
14	20.75607	-6.75607	2.199092
31	29.15734	1.842661	0.116451
35	39.53537	-4.53537	0.520284
47	41.75924	5.24076	0.657712
56	51.89018	4.10982	0.325507
42	37.55861	4.441394	0.525205
2	10.5491	-8.5491	6.928282
4	16.40971	-12.4097	9.384747
7	23.05174	-16.0517	11.1774
17	31.2566	-14.2566	6.502647
28	33.01478	-5.01478	0.761721
61	41.02429	19.97571	9.726655
66	29.69377	36.30623	44.39121
1	3.307286	-2.30729	1.609649
1	5.144667	-4.14467	3.339043
2	7.227033	-5.22703	3.78051
7	9.799366	-2.79937	0.79969
6	10.35058	-4.35058	1.828646
18	12.86167	5.138332	2.052801
23	9.309398	13.6906	20.13369

Decision rule: If $\chi^2_{STAT} > 42.9798$, reject H_0 .

Test statistic: $\chi^2_{STAT} = \sum_{\text{all cells}} \frac{(f_0 - f_e)^2}{f_e} = 229.7554$.

Decision: Because $\chi^2_{STAT} = 229.7554 > 42.9798$ reject H_0 . There is evidence to conclude that there is a relationship between the frequency of Facebook posts and age.

12.26 Because $\chi^2_{STAT} = 6.6876 < 12.5916$, do not reject H_0 . There is insufficient evidence of a relationship between consumer segment and geographic region.

12.28 (a) 31. **(b)** 29. **(c)** 27. **(d)** 25.

12.30 40 and 79.

12.32 (a) The ranks for Sample 1 are 1, 2, 4, 5, and 10. The ranks for Sample 2 are 3, 6.5, 6.5, 8, 9, and 11. **(b)** 22. **(c)** 44.

12.34 Because $T_1 = 22 > 20$, do not reject H_0 .

12.36 (a) The data are ordinal. **(b)** The two-sample t test is inappropriate because the data can only be placed in ranked order. **(c)** Because $Z_{STAT} = -2.2054 < -1.96$, reject H_0 . There is evidence of a significance difference in the median rating of California Cabernets and Washington Cabernets.

12.38 (a) $H_0: M_1 = M_2$, where Populations: 1 = Wing A, 2 = Wing B. $H_1: M_1 \neq M_2$.

Population 1 sample: Sample size 20, sum of ranks 561

Population 2 sample: Sample size 20, sum of ranks 259

$$\begin{aligned}\mu_{T_1} &= \frac{n_1(n+1)}{2} = \frac{20(40+1)}{2} = 410 \\ \sigma_{T_1} &= \sqrt{\frac{n_1 n_2 (n+1)}{12}} = \sqrt{\frac{20(20)(40+1)}{12}} = 36.9685 \\ Z_{STAT} &= \frac{T_1 - \mu_{T_1}}{\sigma_{T_1}} = \frac{561 - 410}{36.9685} = 4.0846\end{aligned}$$

Decision: Because $Z_{STAT} = 4.0846 > 1.96$ (or p -value = 0.0000 < 0.05), reject H_0 . There is sufficient evidence of a difference in the median delivery time in the two wings of the hotel.

(b) The results of (a) are consistent with the results of Problem 10.65.

12.40 (a) Because $Z_{STAT} = 1.5965 < 1.96$, do not reject H_0 . There is insufficient evidence to conclude that there is a difference in the median brand value between the two sectors. **(b)** You must assume approximately equal variability in the two populations. **(c)** Using the pooled-variance t test you rejected the null hypothesis and the separate-variance t test did not allow you to reject the null hypothesis and conclude in Problem 10.17 that the mean brand value is different between the two sectors. In this test, using the Wilcoxon rank sum test with large-sample Z approximation did not allow you to reject the null hypothesis and conclude that the median brand value differs between the two sectors.

12.42 (a) Because $-1.96 < Z_{STAT} = 0.7245 < 1.96$ (or the p -value = 0.4687 > 0.05), do not reject H_0 . There is not enough evidence to conclude that there is a difference in the median battery life between subcompact cameras and compact cameras. **(b)** You must assume approximately equal variability in the two populations. **(c)** Using the pooled-variance t -test, you do not reject the null hypothesis ($t = -2.1199 < -0.6181 < 2.1199$; p -value = 0.5452 > 0.05) and conclude that there is insufficient evidence of a difference in the mean battery life between the two types of digital cameras in Problem 10.11 (a).

12.44 (a) Decision rule: If $H > \chi^2_U = 15.086$, reject H_0 . **(b)** Because $H = 13.77 < 15.086$, do not reject H_0 .

12.46 (a) $H = 13.517 > 7.815$, p -value = 0.0036 < 0.05, reject H_0 . There is sufficient evidence of a difference in the median waiting time in the four locations. **(b)** The results are consistent with those of Problem 11.9.

12.48 (a) $H = 19.3269 > 9.488$, reject H_0 . There is evidence of a difference in the median ratings of the ads. **(b)** The results are consistent

with those of Problem 11.10. **(c)** Because the combined scores are not true continuous variables, the nonparametric Kruskal-Wallis rank test is more appropriate because it does not require that the scores be normally distributed.

12.50 (a) Because $H = 22.0357 > 7.815$ or the p -value is approximately 0, reject H_0 . There is sufficient evidence of a difference in the median cost associated with importing a standardized cargo of goods by sea transport across the four global regions. **(b)** The results are the same.

12.56 (a) Because $\chi^2_{STAT} = 0.412 < 3.841$, do not reject H_0 . There is insufficient evidence to conclude that there is a relationship between a student's gender and pizzeria selection. **(b)** Because $\chi^2_{STAT} = 2.624 < 3.841$, do not reject H_0 . There is insufficient evidence to conclude that there is a relationship between a student's gender and pizzeria selection. **(c)** Because $\chi^2_{STAT} = 4.956 < 5.991$, do not reject H_0 . There is insufficient evidence to conclude that there is a relationship between price and pizzeria selection. **(d)** p -value = 0.0839. The probability of a sample that gives a test statistic equal to or greater than 4.956 is 8.39% if the null hypothesis of no relationship between price and pizzeria selection is true.

12.58 (a) Because $\chi^2_{STAT} = 11.895 < 12.592$, do not reject H_0 . There is not enough evidence to conclude that there is a relationship between the attitudes of employees toward the use of self-managed work teams and employee job classification. **(b)** Because $\chi^2_{STAT} = 3.294 < 12.592$, do not reject H_0 . There is insufficient evidence to conclude that there is a relationship between the attitudes of employees toward vacation time without pay and employee job classification.

CHAPTER 13

13.2 (a) Yes. **(b)** No. **(c)** No. **(d)** Yes.

13.4 (a) The scatter plot shows a positive linear relationship.

(b) For each increase in alcohol percentage of 1.0, mean predicted mean wine quality is estimated to increase by 0.5624.

(c) $\hat{Y} = -0.3529 + 0.5624X = -0.3529 + 0.5624(10) = 5.2715$.

(d) Wine quality appears to be affected by the alcohol percentage. Each increase of 1% in alcohol leads to a mean increase in wine quality of a little more than half a unit.

13.6 (b) $b_0 = -2.37$, $b_1 = 0.0501$. **(c)** For every cubic foot increase in the amount moved, predicted mean labor hours are estimated to increase by 0.0501. **(d)** 22.67 labor hours. **(e)** That as expected, the labor hours are affected by the amount to be moved.

13.8 (b) $b_0 = -601.9291$, $b_1 = 5.9316$. **(c)** For each additional million-dollar increase in revenue, the mean value is predicted to increase by an estimated \$5.9316 million. Literal interpretation of b_0 is not meaningful because an operating franchise cannot have zero revenue. **(d)** \$880.9832 million. **(e)** That the value of the franchise can be expected to increase as revenue increases.

13.10 (b) $b_0 = 4.8445$, $b_1 = 0.1631$. **(c)** For each increase of \$1 million of box office gross, the predicted DVD revenue is estimated to increase by \$0.1631 million. **(d)** $\hat{Y} = b_0 + b_1X$. $\hat{Y} = 4.8445 + 0.1631(100) = \21.1588 million. **(e)** You can conclude that the mean predicted increase in DVD sales is \$163,100 for each million-dollar increase in movie gross.

13.12 $r^2 = 0.90$. 90% of the variation in the dependent variable can be explained by the variation in the independent variable.

13.14 $r^2 = 0.75$. 75% of the variation in the dependent variable can be explained by the variation in the independent variable.

13.16 (a) $r^2 = \frac{SSR}{SST} = \frac{21.8677}{64.0000} = 0.3417$, 34.17% of the variation in wine quality can be explained by the variation in the percentage of alcohol.

$$\text{(b)} S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{42.1323}{48}} = 0.9369.$$

(c) Based on (a) and (b), the model should be somewhat useful for predicting wine quality.

13.18 (a) $r^2 = 0.8892$, 88.92% of the variation in labor hours can be explained by the variation in cubic feet moved. (b) $S_{YX} = 5.0314$.

(c) Based on (a) and (b), the model should be very useful for predicting the labor hours.

13.20 (a) $r^2 = 0.8219$, 82.19% of the variation in the value of a baseball franchise can be explained by the variation in its annual revenue.

(b) $S_{YX} = 165.3106$. (c) Based on (a) and (b), the model should be useful for predicting the value of a baseball franchise.

13.22 (a) $r^2 = 0.5123$, 51.23% of the variation in DVD revenue can be explained by the variation in box office gross. (b) $S_{YX} = 12.2279$.

The variation of DVD revenue around the prediction line is \$12.2279 million. The typical difference between actual DVD revenue and the predicted DVD revenue using the regression equation is approximately \$12.2279 million. (c) Based on (a) and (b), the model may not be useful for predicting DVD revenue. (d) Other variables that might explain the variation in DVD revenue could be the amount spent on advertising, the timing of the release of the DVDs, and the type of movie.

13.24 A residual analysis of the data indicates a pattern, with sizable clusters of consecutive residuals that are either all positive or all negative. This pattern indicates a violation of the assumption of linearity. A curvilinear model should be investigated.

13.26 There does not appear to be a pattern in the residual plot. The assumptions of regression do not appear to be seriously violated.

13.28 Based on the residual plot, there does not appear to be a curvilinear pattern in the residuals. The assumptions of normality and equal variance do not appear to be seriously violated.

13.30 Based on the residual plot, there appears to be an outlier in the residuals, but no evidence of a pattern.

13.32 (a) An increasing linear relationship exists. (b) There is evidence of a strong positive autocorrelation among the residuals.

13.34 (a) No, because the data were not collected over time. (b) If data were collected at a single store had been selected and studied over a period of time, you would compute the Durbin-Watson statistic.

13.36 (a)

$$b_1 = \frac{SSXY}{SSX} = \frac{201,399.05}{12,495.626} = 0.0161$$

$$b_0 = \bar{Y} - b_1 \bar{X} = 71.2621 - 0.0161(4,393) = 0.458.$$

(b) $\hat{Y} = 0.458 + 0.0161X = 0.458 + 0.0161(4,500) = 72.908$, or \$72,908. (c) There is no evidence of a pattern in the residuals over time.

$$\text{(d)} D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} = \frac{1,243.2244}{599.0683} = 2.08 > 1.45. \text{ There is no evidence of positive autocorrelation among the residuals.}$$

(e) Based on a residual analysis, the model appears to be adequate.

13.38 (a) $b_0 = -2.535$, $b_1 = 0.06073$. (b) \$2,505.40. (d) D = 1.64 > d_U = 1.42, so there is no evidence of positive autocorrelation among the residuals. (e) The plot shows some nonlinear pattern, suggesting that a nonlinear model might be better. Otherwise, the model appears to be adequate.

13.40 (a) 3.00. (b) ± 2.1199 . (c) Reject H_0 . There is evidence that the fitted linear regression model is useful. (d) $1.32 \leq \beta_1 \leq 7.68$.

$$\text{(e)} b_1 \pm t_{\alpha/2} S_{b_1} = 0.5624 \pm 2.0106(0.1127) 0.3359 \leq \beta_1 \leq 0.7890.$$

13.44 (a) $t_{STAT} = 16.52 > 2.0322$; reject H_0 . There is evidence of a linear relationship between the number of cubic feet moved and labor hours. (b) $0.0439 \leq \beta_1 \leq 0.0562$.

13.46 (a) $t_{STAT} = 11.3668 > 2.0484$ or because the p -value is 0.0000, reject H_0 at the 5% level of significance. There is evidence of a linear relationship between annual revenue and franchise value. (b) $4.8627 \leq \beta_1 \leq 7.0006$.

13.48 (a) $t_{STAT} = 6.5626 > 2.0195$ or because the p -value = 0.0000 < 0.05; reject H_0 . There is evidence of a linear relationship between box office gross and sales of DVDs. (b) $0.1129 \leq \beta_1 \leq 0.2133$.

13.50 (a) (% daily change in MDU) = $b_0 + 3.0$ (% daily change in S&P Midcap 400 index). (b) If the S&P Midcap 400 gains 10% in a year, MDU is expected to gain an estimated 30%. (c) If the S&P Midcap 400 loses 20% in a year, MDU is expected to lose an estimated 60%. (d) Risk takers will be attracted to leveraged funds, and risk-averse investors will stay away.

13.52 (a), (b) First weekend and U.S. gross: $r = 0.7264$, $t_{STAT} = 2.5893 > 2.4469$, p -value = 0.0413 < 0.05. reject H_0 . At the 0.05 level of significance, there is evidence of a linear relationship between first weekend sales and U.S. gross. First weekend and worldwide gross: $r = 0.8234$, $t_{STAT} = 3.5549 > 2.4469$, p -value = 0.0120 < 0.05. reject H_0 . At the 0.05 level of significance, there is evidence of a linear relationship between first weekend sales and worldwide gross. U.S. gross and worldwide gross: $r = 0.9629$, $t_{STAT} = 8.7456 > 2.4469$, p -value = 0.0001 < 0.05. Reject H_0 . At the 0.05 level of significance, there is evidence of a linear relationship between U.S. gross and worldwide gross.

13.54 (a) $r = 0.7042$. There appears to be a moderate positive linear relationship between social media networking and the GDP per capita. (b) $t_{STAT} = 4.3227$, p -value = 0.0004 < 0.05. Reject H_0 . At the 0.05 level of significance, there is a significant linear relationship between social media networking and the GDP per capita. (c) There appears to be a strong relationship.

13.56 (a) $15.95 \leq \mu_{Y|X=4} \leq 18.05$. (b) $14.651 \leq Y_{X=4} \leq 19.349$.

$$\text{(c)} \hat{Y} = -0.3529 + (0.5624)(10) = 5.2715 \hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{h_i} = 5.2715 \pm 2.0106(0.9369) \sqrt{0.0249} 4.9741 \leq \mu_{Y|X=10} \leq 5.5690.$$

$$\text{(d)} \hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{1 + h_i} = 5.2715 \pm 2.0106(9.369) \sqrt{1 + 0.0249} 3.3645 \leq Y_{X=10} \leq 7.1786.$$

(e) Part (b) provides a prediction interval for the individual response given a specific value of the independent variable, and part (a) provides

an interval estimate for the mean value, given a specific value of the independent variable. Because there is much more variation in predicting an individual value than in estimating a mean value, a prediction interval is wider than a confidence interval estimate.

- 13.60** (a) $20.799 \leq \mu_{Y|X=500} \leq 24.542$. (b) $12.276 \leq Y_{X=500} \leq 33.065$. (c) You can estimate a mean more precisely than you can predict a single observation.

13.62 (a) $814.3841 \leq \mu_{Y|X=250} \leq 947.5823$. (b) $535.8727 \leq Y_{X=250} \leq 1,226.094$. (c) Part (b) provides a prediction interval for an individual response given a specific value of X , and part (a) provides a confidence interval estimate for the mean value, given a specific value of X . Because there is much more variation in predicting an individual value than in estimating a mean, the prediction interval is wider than the confidence interval.

- 13.74** (a) $b_0 = 24.84$, $b_1 = 0.14$. (b) For each additional case, the predicted delivery time is estimated to increase by 0.14 minute. (c) 45.84. (d) No, 500 is outside the relevant range of the data used to fit the regression equation. (e) $r^2 = 0.972$. (f) There is no obvious pattern in the residuals, so the assumptions of regression are met. The model appears to be adequate. (g) $t_{STAT} = 24.88 > 2.1009$; reject H_0 . (h) $44.88 \leq \mu_{Y|X=150} \leq 46.80$. $41.56 \leq Y_{X=150} \leq 50.12$. (i) The number of cases explains almost all of the variation in delivery time.

13.76 (a) $b_0 = 276.848$, $b_1 = 50.8031$. (b) For each additional 1,000 square feet in the size of the house, the mean assessed value is predicted to increase by \$50,803.10. The estimated selling price of a house with a 0 size is \$276,848 thousand. However, this interpretation is not meaningful because the size of the house cannot be 0. (c) $\hat{Y} = 276.848 + 50.8031(2) = 378.4542$ thousand dollars. (d) $r^2 = 0.3273$. So 32.73% of the variation in assessed value be explained by the variation in size. (e) Neither the residual plot nor the normal probability plot reveals any potential violation of the linearity, equal variance, and normality assumptions. (f) $t_{STAT} = 3.6913 > 2.0484$, p -value is 0.0009. Because p -value < 0.05 , reject H_0 . There is evidence of a linear relationship between assessed value and size. (g) $22.6113 \leq \beta_1 \leq 78.9949$. (h) The size of the house is somewhat useful in predicting the assessed value, but since only 32.73% of the variation in assessed value is explained by variation in size, other variables should be considered.

13.78 (a) $b_0 = 0.30$, $b_1 = 0.00487$. (b) For each additional point on the GMAT score, the predicted GPA is estimated to increase by 0.00487. Because a GMAT score of 0 is not possible, the Y intercept does not have a practical interpretation. (c) 3.222. (d) $r^2 = 0.798$. (e) There is no obvious pattern in the residuals, so the assumptions of regression are met. The model appears to be adequate. (f) $t_{STAT} = 8.43 > 2.1009$; reject H_0 . (g) $3.144 \leq \mu_{Y|X=600} \leq 3.301$, $2.866 \leq Y_{X=600} \leq 3.559$. (h) $.00366 \leq \beta_1 \leq .00608$. (i) Most of the variation in GPA can be explained by variation in the GMAT score.

13.80 (a) There is no clear relationship shown on the scatter plot. (c) Looking at all 23 flights, when the temperature is lower, there is likely to be some O-ring damage, particularly if the temperature is below 60 degrees. (d) 31 degrees is outside the relevant range, so a prediction should not be made. (e) Predicted $Y = 18.036 - 0.240X$, where X = temperature and Y = O-ring damage. (g) A nonlinear model would be more appropriate. (h) The appearance on the residual plot of a nonlinear pattern indicates that a nonlinear model would be better. It also appears that the normality assumption is invalid.

13.82 (a) $b_0 = -132.5214$, $b_1 = 5.2286$. (b) For each additional million-dollar increase in revenue, the franchise value will increase by an estimated \$5.2286 million. Literal interpretation of b_0 is not meaningful

because an operating franchise cannot have zero revenue.

- (c) \$651.7731 million. (d) $r^2 = 0.836$. 83.6% of the variation in the value of an NBA franchise can be explained by the variation in its annual revenue. (e) There does not appear to be a pattern in the residual plot. The assumptions of regression do not appear to be seriously violated. (f) $t_{STAT} = 11.9452 > 2.0484$ or because the p -value is 0.0000, reject H_0 at the 5% level of significance. There is evidence of a linear relationship between annual revenue and franchise value. (g) $613.6103 \leq \mu_{Y|X=150} \leq 689.9359$. (h) $486.9282 \leq Y_{X=150} \leq 816.618$. (i) The strength of the relationship between revenue and value is about the same for NBA franchises, European soccer teams, and Major League Baseball teams.

- 13.84** (a) $b_0 = -2,629.222$, $b_1 = 82.472$. (b) For each additional centimeter in circumference, the weight is estimated to increase by 82.472 grams. (c) 2,319.08 grams. (d) Yes, since circumference is a very strong predictor of weight. (e) $r^2 = 0.937$. (f) There appears to be a nonlinear relationship between circumference and weight. (g) p -value is virtually $0 < 0.05$; reject H_0 . (h) $72.7875 \leq \beta_1 \leq 92.156$.

13.86 (a) The correlation between compensation and stock performance is 0.1719. (b) $t_{STAT} = 2.2615 > 1.9742$; p -value = 0.025 < 0.05 . The correlation between compensation and stock performance is significant, but only 2.95% of the variation in compensation can be explained by return. (c) The small correlation between compensation and stock performance was surprising (or maybe it shouldn't have been!).

CHAPTER 14

14.2 (a) For each one-unit increase in X_1 , you estimate that the mean of Y will decrease 2 units, holding X_2 constant. For each one-unit increase in X_2 , you estimate that the mean of Y will increase 7 units, holding X_1 constant. (b) The Y intercept, equal to 50, estimates the value of Y when both X_1 and X_2 are 0.

- 14.4** (a) $\hat{Y} = -2.72825 + 0.047114X_1 + 0.011947X_2$. (b) For a given number of orders, for each increase of \$1,000 in sales, the mean distribution cost is estimated to increase by \$47.114. For a given amount of sales, for each increase of one order, the mean distribution cost is estimated to increase by \$11.95. (c) The interpretation of b_0 has no practical meaning here because it would represent the estimated distribution cost when there were no sales and no orders. (d) $\hat{Y} = -2.72825 + 0.047114(400) + 0.011947(4500) = 69.878$, or \$69,878. (e) $\$66,419.93 \leq \mu_{Y|X} \leq \$73,337.01$. (f) $\$59,380.61 \leq Y_X \leq \$80,376.33$ (g) The interval in (e) is narrower because it is estimating the mean value, not an individual value. (h) The model uses both the number of orders and the amount of sales to predict warehouse distribution cost. This may produce a better model than if only one of these independent variables is included.

14.6 (a) $\hat{Y} = 156.4 + 13.081X_1 + 16.795X_2$. (b) For a given amount of newspaper advertising, each increase by \$1,000 in radio advertising is estimated to result in an increase in mean sales of \$13,081. For a given amount of radio advertising, each increase by \$1,000 in newspaper advertising is estimated to result in an increase in mean sales of \$16,795. (c) When there is no money spent on radio advertising and newspaper advertising, the estimated mean sales is \$156,430.44. (d) Holding the other independent variable constant, newspaper advertising seems to be more effective because its slope is greater.

- 14.8** (a) $\hat{Y} = 532.2883 + 407.1346X_1 - 2.8257X_2$, where X_1 = land area, X_2 = age. (b) For a given age, each increase by one acre in land area is estimated to result in an increase in the mean fair market value by \$407.1346 thousands. For a given land area, each

increase of one year in age is estimated to result in a decrease in the mean fair market value by \$2.8257 thousands.

(c) The interpretation of b_0 has no practical meaning here because it would represent the estimated fair market value of a new house that has no land area. (d) $\hat{Y} = 5,332.2883 + 407.1346(0.25) - 2.8257(55) = \478.6577 thousands. (e) $446.8367 \leq \mu_{Y|X} \leq 510.4788$. (f) $307.2577 \leq Y_X \leq 650.0577$.

14.10 (a) $MSR = 15$, $MSE = 12$. (b) 1.25. (c) $F_{STAT} = 1.25 < 4.10$; do not reject H_0 . (d) 0.20. (e) 0.04.

14.12 (a) $F_{STAT} = 97.69 > 3.89$. Reject H_0 . There is evidence of a significant linear relationship with at least one of the independent variables. (b) $p\text{-value} = 0.0001$. (c) $r^2 = 0.9421$. 94.21% of the variation in the long-term ability to absorb shock can be explained by variation in forefoot-absorbing capability and variation in midsole impact. (d) $r_{adj}^2 = 0.935$.

14.14 (a) $F_{STAT} = 74.13 > 3.467$; reject H_0 . (b) $p\text{-value} = 0$. (c) $r^2 = 0.8759$. 87.59% of the variation in distribution cost can be explained by variation in sales and variation in number of orders. (d) $r_{adj}^2 = 0.8641$.

14.16 (a) $F_{STAT} = 40.16 > 3.522$. Reject H_0 . There is evidence of a significant linear relationship. (b) $p\text{-value} < 0.001$. (c) $r^2 = 0.8087$. 80.87% of the variation in sales can be explained by variation in radio advertising and variation in newspaper advertising. (d) $r_{adj}^2 = 0.7886$.

14.18 (a)–(e) Based on a residual analysis, there is no evidence of a violation of the assumptions of regression. (f) $D = 2.26$ (g) $D = 2.26 > 1.55$. There is no evidence of positive autocorrelation in the residuals.

14.20 (a) There appears to be a quadratic relationship in the plot of the residuals against both radio and newspaper advertising. (b) Since the data are not collected over time, the Durbin-Watson test is not appropriate. (c) Curvilinear terms for both of these explanatory variables should be considered for inclusion in the model.

14.22 (a) The residual analysis reveals no patterns. (b) Since the data are not collected over time, the Durbin-Watson test is not appropriate. (c) There are no apparent violations in the assumptions.

14.24 (a) Variable X_2 has a larger slope in terms of the t statistic of 3.75 than variable X_1 , which has a smaller slope in terms of the t statistic of 3.33. (b) $1.46824 \leq \beta_1 \leq 6.53176$. (c) For $X_1: t_{STAT} = 4/1.2 = 3.33 > 2.1098$, with 17 degrees of freedom for $\alpha = 0.05$. Reject H_0 . There is evidence that X_1 contributes to a model already containing X_2 . For $X_2: t_{STAT} = 3/0.8 = 3.75 > 2.1098$, with 17 degrees of freedom for $\alpha = 0.05$. Reject H_0 . There is evidence that X_2 contributes to a model already containing X_1 . Both X_1 and X_2 should be included in the model.

14.26 (a) 95% confidence interval on $\beta_1: b_1 \pm tS_{b_1}$, $0.0471 \pm 2.0796 (0.0203)$, $0.0049 \leq \beta_1 \leq 0.0893$. (b) For $X_1: t_{STAT} = b_1/S_{b_1} = 0.0471/0.0203 = 2.32 > 2.0796$. Reject H_0 . There is evidence that X_1 contributes to a model already containing X_2 . For $X_2: t_{STAT} = b_1/S_{b_1} = 0.0112/0.0023 = 5.31 > 2.0796$. Reject H_0 . There is evidence that X_2 contributes to a model already containing X_1 . Both X_1 (sales) and X_2 (orders) should be included in the model.

14.28 (a) $9.398 \leq \beta_1 \leq 16.763$. (b) For $X_1: t_{STAT} = 7.43 > 2.093$. Reject H_0 . There is evidence that X_1 contributes to a model already containing X_2 . For $X_2: t_{STAT} = 5.67 > 2.093$. Reject H_0 . There is evidence that X_2 contributes to a model already containing X_1 . Both X_1

(radio advertising) and X_2 (newspaper advertising) should be included in the model.

14.30 (a) $274.1702 \leq \beta_1 \leq 540.0990$. (b) For $X_1: t_{STAT} = 6.2827$ and $p\text{-value} = 0.0000$. Because $p\text{-value} < 0.05$, reject H_0 . There is evidence that X_1 contributes to a model already containing X_2 . For $X_2: t_{STAT} = -4.1475$ and $p\text{-value} = 0.0003$. Because $p\text{-value} < 0.05$ reject H_0 . There is evidence that X_2 contributes to a model already containing X_1 : $F_{STAT} = 30.4533$ $p\text{-value} = 0.0000$. Both X_1 (land area) and X_2 (age) should be included in the model.

14.32 (a) For $X_1: F_{STAT} = 1.25 < 4.96$; do not reject H_0 . For $X_2: F_{STAT} = 0.833 < 4.96$; do not reject H_0 . (b) 0.1111, 0.0769.

14.34 (a) For $X_1: SSR(X_1|X_2) = SSR(X_1 \text{ and } X_2) - SSR(X_2) = 3,368.087 - 3,246.062 = 122.025$, $F_{STAT} = \frac{SSR(X_1|X_2)}{MSE} = \frac{122.025}{477.043/21} = 5.37 > 4.325$. Reject H_0 . There is evidence that X_1 contributes to a model already containing X_2 . For $X_2: SSR(X_2|X_1) = SSR(X_1 \text{ and } X_2) - SSR(X_1) = 3,368.087 - 2,726.822 = 641.265$, $F_{STAT} = \frac{SSR(X_2|X_1)}{MSE} = \frac{641.265}{477.043/21} = 28.23 > 4.325$.

Reject H_0 . There is evidence that X_2 contributes to a model already containing X_1 . Because both X_1 and X_2 make a significant contribution to the model in the presence of the other variable, both variables should be included in the model.

$$(b) r_{Y1,2}^2 = \frac{SSR(X_1|X_2)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_1|X_2)} = \frac{122.025}{3,845.13 - 3,368.087 + 122.025} = 0.2037.$$

Holding constant the effect of the number of orders, 20.37% of the variation in distribution cost can be explained by the variation in sales.

$$r_{Y2,1}^2 = \frac{SSR(X_2|X_1)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_2|X_1)} = \frac{641.265}{3,845.13 - 3,368.087 + 641.265} = 0.5734$$

Holding constant the effect of sales, 57.34% of the variation in distribution cost can be explained by the variation in the number of orders.

14.36 (a) For $X_1: F_{STAT} = 55.28 > 4.381$. Reject H_0 . There is evidence that X_1 contributes to a model containing X_2 . For $X_2: F_{STAT} = 32.12 > 4.381$. Reject H_0 . There is evidence that X_2 contributes to a model already containing X_1 . Because both X_1 and X_2 make a significant contribution to the model in the presence of the other variable, both variables should be included in the model. (b) $r_{Y1,2}^2 = 0.7442$. Holding constant the effect of newspaper advertising, 74.42% of the variation in sales can be explained by the variation in radio advertising. $r_{Y2,1}^2 = 0.6283$. Holding constant the effect of radio advertising, 62.83% of the variation in sales can be explained by the variation in newspaper advertising.

14.38 (a) Holding constant the effect of X_2 , for each increase of one unit of X_1 , Y increases by 4 units. (b) Holding constant the effect of X_1 , for each increase of one unit of X_2 , Y increases by 2 units. (c) Because $t_{STAT} = 3.27 > 2.1098$, reject H_0 . Variable X_2 makes a significant contribution to the model.

14.40 (a) $\hat{Y} = 243.7371 + 9.2189X_1 + 12.6967X_2$, where X_1 = number of rooms and X_2 = neighborhood (east = 0). **(b)** Holding constant the effect of neighborhood, for each additional room, the mean selling price is estimated to increase by 9.2189 thousands of dollars, or \$9,218.9. For a given number of rooms, a west neighborhood is estimated to increase the mean selling price over an east neighborhood by 12.6967 thousands of dollars, or \$12,696.7.

(c) $\hat{Y} = 243.7371 + 9.2189(9) + 12.6967(0) = 326.7076$, or \$326,707.6. $\$309,560.04 \leq Y_X \leq 343,855.1$. $\$321,471.44 \leq \mu_{Y|X} \leq \$331,943.71$. **(d)** Based on a residual analysis, the model appears to be adequate. **(e)** $F_{STAT} = 55.39$, the p -value is virtually 0. Because p -value < 0.05, reject H_0 . There is evidence of a significant relationship between selling price and the two independent variables (rooms and neighborhood). **(f)** For X_1 : $t_{STAT} = 8.9537$, the p -value is virtually 0. Reject H_0 . Number of rooms makes a significant contribution and should be included in the model. For X_2 : $t_{STAT} = 3.5913$, p -value = 0.0023 < 0.05. Reject H_0 . Neighborhood makes a significant contribution and should be included in the model. Based on these results, the regression model with the two independent variables should be used. **(g)** $7.0466 \leq \beta_1 \leq 11.3913$.

(h) $5.2378 \leq \beta_2 \leq 20.1557$. **(i)** $r_{adj}^2 = 0.851$. **(j)** $r_{Y1,2}^2 = 0.825$. Holding constant the effect of neighborhood, 82.5% of the variation in selling price can be explained by variation in number of rooms. $r_{Y2,1}^2 = 0.431$. Holding constant the effect of number of rooms, 43.1% of the variation in selling price can be explained by variation in neighborhood.

(k) The slope of selling price with number of rooms is the same, regardless of whether the house is located in an east or west neighborhood. **(l)** $\hat{Y} = 253.95 + 8.032X_1 - 5.90X_2 + 2.089X_1X_2$. For $X_1 X_2$, p -value = 0.330. Do not reject H_0 . There is no evidence that the interaction term makes a contribution to the model. **(m)** The model in (b) should be used. **(n)** The number of rooms and the neighborhood both significantly affect the selling price, but the number of rooms has a greater effect.

14.42 (a) Predicted time = $8.01 + 0.00523$ Depth – 2.105 Dry. **(b)** Holding constant the effect of type of drilling, for each foot increase in depth of the hole, the mean drilling time is estimated to increase by 0.00523 minutes. For a given depth, a dry drilling hole is estimated to reduce the drilling time over wet drilling by a mean of 2.1052 minutes. **(c)** 6.428 minutes, $6.210 \leq \mu_{Y|X} \leq 6.646$, $4.923 \leq Y_X \leq 7.932$. **(d)** The model appears to be adequate. **(e)** $F_{STAT} = 111.11 > 3.09$; reject H_0 . **(f)** $t_{STAT} = 5.03 > 1.9847$; reject H_0 . $t_{STAT} = -14.03 < -1.9847$; reject H_0 . Include both variables. **(g)** $0.0032 \leq \beta_1 \leq 0.0073$. **(h)** $-2.403 \leq \beta_2 \leq -1.808$. **(i)** 69.0%. **(j)** 0.207, 0.670. **(k)** The slope of the additional drilling time with the depth of the hole is the same, regardless of the type of drilling method used. **(l)** The p -value of the interaction term = 0.462 > 0.05, so the term is not significant and should not be included in the model. **(m)** The model in part (b) should be used. Both variables affect the drilling time. Dry drilling holes should be used to reduce the drilling time.

14.44 (a) $\hat{Y} = 31.5594 + 0.0296X_1 + 0.0041X_2 + 0.000017159X_1X_2$, where X_1 = sales, X_2 = orders, p -value = 0.3249 > 0.05. Do not reject H_0 . There is not enough evidence that the interaction term makes a contribution to the model. **(b)** Because there is insufficient evidence of any interaction effect between sales and orders, the model in Problem 14.4 should be used.

14.46 (a) The p -value of the interaction term = 0.002 < 0.05, so the term is significant and should be included in the model. **(b)** Use the model developed in this problem.

14.48 (a) For $X_1 X_2$, p -value = 0.2353 > 0.05. Do not reject H_0 . There is insufficient evidence that the interaction term makes a contribution to the model. **(b)** Because there is not enough evidence of an interaction effect

between total staff present and remote hours, the model in Problem 14.7 should be used.

14.50 Holding constant the effect of other variables, the natural logarithm of the estimated odds ratio for the dependent categorical response will increase by 2.2 for each unit increase in the particular independent variable.

14.52 0.4286.

14.54 (a) $\ln(\text{estimated odds ratio}) = -6.9394 + 0.1395X_1 + 2.7743X_2 = -6.9394 + 0.1395(36) + 2.7743(0) = -1.91908$. Estimated odds ratio = $e^{-1.91908} = 0.1470$. Estimated Probability of Success = Odds Ratio/(1 + Odds Ratio) = $0.1470/(1 + 0.1470) = 0.1260$. **(b)** From the text discussion of the example, 70.2% of the individuals who charge \$36,000 per annum and possess additional cards can be expected to purchase the premium card. Only 12.60% of the individuals who charge \$36,000 per annum and do not possess additional cards can be expected to purchase the premium card. For a given amount of money charged per annum, the likelihood of purchasing a premium card is substantially higher among individuals who already possess additional cards than for those who do not possess additional cards. **(c)** $\ln(\text{estimated odds ratio}) = -6.9394 + 0.13957X_1 + 2.7743X_2 = -6.9394 + 0.1395(18) + 2.7743(0) = -4.4298$. Estimated odds ratio = $e^{-4.4298} = 0.0119$. Estimated Probability of Success = Odds Ratio/(1 + Odds Ratio) = $0.0119/(1 + 0.0119) = 0.01178$. **(d)** Among individuals who do not purchase additional cards, the likelihood of purchasing a premium card diminishes dramatically with a substantial decrease in the amount charged per annum.

14.56 Using Microsoft Excel **(a)** $\ln(\text{estimated odds}) = -47.4723 + 1.3099$ fixed acidity + 90.5722 chlorides + 9.777 pH. **(b)** Holding constant the effect of chlorides and pH, for each increase of one point in fixed acidity, $\ln(\text{estimated odds})$ increases by an estimate of 1.3099. Holding constant the effect of fixed acidity and pH, for each increase of one point in chlorides, $\ln(\text{estimated odds})$ increases by an estimate of 90.5722. Holding constant the effect of fixed acidity and chlorides, for each increase of one point in pH, $\ln(\text{estimated odds})$ increases by an estimate of 9.777. **(c)** 0.3686. **(d)** Deviance = 54.456, p -value = 1.0000, do not reject H_0 , so model is adequate. **(e)** For fixed acidity: $Z_{STAT} = 3.17 > 1.96$, reject H_0 . For chlorides: $Z_{STAT} = 4.00 > 1.96$, reject H_0 . For pH: $Z_{STAT} = 2.9738 > 1.96$, reject H_0 . Each variable makes a significant contribution to the model. **(f)** Fixed acidity, chlorides, and pH are all important factors in distinguishing between white and red wines.

14.58 (a) Using Microsoft Excel **(a)** $\ln(\text{estimated odds}) = -0.6048 + 0.0938$ claims/year + 1.8108 new business **(b)** Holding constant the effects of whether the policy is new, for each increase of the number of claims submitted per year by the policy holder, $\ln(\text{odds})$ increases by an estimate of 0.0938. Holding constant the number of claims submitted per year by the policy holder, $\ln(\text{odds})$ is estimated to be 1.8108 higher when the policy is new as compared to when the policy is not new. **(c)** $\ln(\text{estimated odds ratio}) = -0.6048 + 0.0938(1) + 1.8108(1) = 1.2998$. Estimated odds ratio = $e^{1.2998} = 3.6684$. Estimated Probability of the Event of Interest = Estimated Odds Ratio/(1 + Estimated Odds Ratio) = 0.7858 **(d)** The deviance statistic is 119.4353 with a χ^2 distribution of 95 d.f. and p -value = 0.0457 < 0.05. Reject H_0 . The model is not a good fitting model. **(e)** For claims/year: $Z_{STAT} = 0.1865$, p -value = 0.8521 > 0.05. Do not reject H_0 . There is insufficient evidence that the number of claims submitted per year by the policy holder makes a significant contribution to the logistic regression model. For new business: $Z_{STAT} = 2.2261$, p -value = 0.0260 < 0.05. Reject H_0 .

There is sufficient evidence that whether the policy is new makes a significant contribution to the logistic model regression. (f) $\ln(\text{estimated odds}) = -1.0125 + 0.9927 \text{ claims/year}$ (g) $\ln(\text{estimated odds}) = -0.5423 + 1.9286 \text{ new business}$ (h) The deviance statistic for (f) is 125.0102 with a χ^2 distribution of 96 d.f. and $p\text{-value} = 0.0250 < 0.05$. Reject H_0 . The model is not a good fitting model. The deviance statistic for (g) is 119.4702 with a χ^2 distribution of 96 d.f. and $p\text{-value} = 0.0526 > 0.05$. Do not reject H_0 . The model is a good fitting model. The model in (g) should be used to predict a fraudulent claim.

14.60 Largest Cook's $D = 0.5637 < 0.8149$ so no need for deletion of any cases.

14.62 Largest Cook's $D = 0.652 < 0.8177$ for observation 2. h_i for this observation $= 0.2924 > 0.2727$ and $t_i = 2.4431 > 1.7291$, so you may wish to delete observation 2 and determine if that affects the fit of the model.

14.64 Largest Cook's $D = 0.5344 < 0.8149$ so no need for deletion of any cases.

14.76 (a) $\hat{Y} = -3.9152 + 0.0319X_1 + 4.2228X_2$, where $X_1 = \text{number cubic feet moved}$ and $X_2 = \text{number of pieces of large furniture}$. (b) Holding constant the number of pieces of large furniture, for each additional cubic foot moved, the mean labor hours are estimated to increase by 0.0319. Holding constant the amount of cubic feet moved, for each additional piece of large furniture, the mean labor hours are estimated to increase by 4.2228. (c) $\hat{Y} = -3.9152 + 0.0319(500) + 4.2228(2) = 20.4926$. (d) Based on a residual analysis, the errors appear to be normally distributed. The equal-variance assumption might be violated because the variances appear to be larger around the center region of both independent variables. There might also be violation of the linearity assumption. A model with quadratic terms for both independent variables might be fitted. (e) $F_{STAT} = 228.80$, $p\text{-value}$ is virtually 0. Because $p\text{-value} < 0.05$, reject H_0 . There is evidence of a significant relationship between labor hours and the two independent variables (the amount of cubic feet moved and the number of pieces of large furniture). (f) The $p\text{-value}$ is virtually 0. The probability of obtaining a test statistic of 228.80 or greater is virtually 0 if there is no significant relationship between labor hours and the two independent variables (the amount of cubic feet moved and the number of pieces of large furniture). (g) $r^2 = 0.9327$. 93.27% of the variation in labor hours can be explained by variation in the number of cubic feet moved and the number of pieces of large furniture. (h) $r_{adj}^2 = 0.9287$. (i) For X_1 : $t_{STAT} = 6.9339$, the $p\text{-value}$ is virtually 0. Reject H_0 . The number of cubic feet moved makes a significant contribution and should be included in the model. For X_2 : $t_{STAT} = 4.6192$, the $p\text{-value}$ is virtually 0. Reject H_0 . The number of pieces of large furniture makes a significant contribution and should be included in the model. Based on these results, the regression model with the two independent variables should be used. (j) For X_1 : $t_{STAT} = 6.9339$, the $p\text{-value}$ is virtually 0. The probability of obtaining a sample that will yield a test statistic farther away than 6.9339 is virtually 0 if the number of cubic feet moved does not make a significant contribution, holding the effect of the number of pieces of large furniture constant. For X_2 : $t_{STAT} = 4.6192$, the $p\text{-value}$ is virtually 0. The probability of obtaining a sample that will yield a test statistic farther away than 4.6192 is virtually 0 if the number of pieces of large furniture does not make a significant contribution, holding the effect of the amount of cubic feet moved constant. (k) $0.0226 \leq \beta_1 \leq 0.0413$. You are 95% confident that the mean labor hours will increase by between 0.0226 and 0.0413 for each additional cubic foot moved, holding constant the number of pieces of large furniture. In Problem 13.44, you are 95% confident that the labor hours will increase by between 0.0439 and 0.0562 for each additional cubic foot moved, regardless of the number of pieces of large

furniture. (l) $r_{Y1,2}^2 = 0.5930$. Holding constant the effect of the number of pieces of large furniture, 59.3% of the variation in labor hours can be explained by variation in the amount of cubic feet moved. $r_{Y2,1}^2 = 0.3927$. Holding constant the effect of the number of cubic feet moved, 39.27% of the variation in labor hours can be explained by variation in the number of pieces of large furniture. (m) Both the number of cubic feet moved and the number of large pieces of furniture are useful in predicting the labor hours, but the cubic feet removed is more important.

14.78 (a) $\hat{Y} = 257.9033 + 53.3606X_1 + 0.2521X_2$, where $X_1 = \text{house size}$ and $X_2 = \text{age}$. (b) Holding constant the age, for each additional thousand square feet in the size of the house, the mean assessed value is estimated to increase by 53.3606 thousand dollars. Holding constant the size of the house, for each additional year in age, the assessed value is estimated to increase by 0.2521 thousand dollars. (c) $\hat{Y} = 257.9033 + 53.3606(2) + 0.2521(55) = 378.4093$ thousand dollars. (d) Based on a residual analysis, the model appears to be adequate. (e) $F_{STAT} = 6.6459$, the $p\text{-value} = 0.0045$. Because $p\text{-value} < 0.05$, reject H_0 . There is evidence of a significant relationship between assessed value and the two independent variables (size of the house and age). (f) The $p\text{-value}$ is 0.0045. The probability of obtaining a test statistic of 6.6459 or greater is virtually 0 if there is no significant relationship between assessed value and the two independent variables (size of the house and age). (g) $r^2 = 0.3299$. 32.99% of the variation in assessed value can be explained by variation in the size of the house and age. (h) $r_{adj}^2 = 0.2803$. (i) For X_1 : $t_{STAT} = 3.3128$, the $p\text{-value}$ is 0.0026. Reject H_0 . The size of the house makes a significant contribution and should be included in the model. For X_2 : $t_{STAT} = 0.3203$, $p\text{-value} = 0.7512 > 0.05$. Do not reject H_0 . Age does not make a significant contribution and should not be included in the model. Based on these results, the regression model with only the size of the house should be used. (j) For X_1 : $t_{STAT} = 3.3128$, the $p\text{-value}$ is virtually 0. The probability of obtaining a sample that will yield a test statistic farther away than 3.3128 is 0.0026 if the house size does not make a significant contribution, holding age constant. For X_2 : $t_{STAT} = 0.3203$, the $p\text{-value}$ is 0.7512. The probability of obtaining a sample that will yield a test statistic farther away than 0.3203 is 0.7512 if the age does not make a significant contribution holding the effect of the house size constant. (k) $20.3109 \leq \beta_1 \leq 86.4104$. You are 95% confident that the assessed value will increase by an amount somewhere between \$20.3109 thousand and \$86.4104 thousand for each additional thousand square foot increase in house size, holding constant the age of the house. In Problem 13.76, you are 95% confident that the assessed value will increase by an amount somewhere between \$22.6113 thousand and \$78.9949 thousand for each additional 1,000 square foot increase in house size, regardless of the age of the house. (l) $r_{Y1,2}^2 = 0.2890$. Holding constant the effect of the age of the house, 28.9% of the variation in assessed value can be explained by variation in the size of the house. $r_{Y2,1}^2 = 0.0038$. Holding constant the effect of the size of the house, 0.38% of the variation in assessed value can be explained by variation in the age of the house.

14.80 (a) $\hat{Y} = 694.9557 + 8.6059X_1 + 2069X_2$, where $X_1 = \text{assessed value}$ and $X_2 = \text{age}$. (b) Holding age constant, for each additional \$1,000, the taxes are estimated to increase by a mean of \$8.61 thousand. Holding assessed value constant, for each additional year, the taxes are estimated to increase by \$2.069. (c) $\hat{Y} = 694.9557 + 8.6059(400) + 2.069(50) = 4,240.542$ dollars. (d) Based on a residual analysis, the errors appear to be normally distributed. The equal-variance assumption appears to be valid. (e) $F_{STAT} = 22.0699$, $p\text{-value} = 0.0000$. Because $p\text{-value} = 0.0000 < 0.05$, reject H_0 . There is evidence of a significant relationship between taxes and the two independent variables (assessed value and age). (f) $p\text{-value} = 0.0000$. The probability of obtaining an F_{STAT} test statistic of 22.0699 or greater is virtually 0 if there is no

significant relationship between taxes and the two independent variables (assessed value and age). (g) $r^2 = 0.6205$. 62.05% of the variation in taxes can be explained by variation in assessed value and age.

(h) $r_{adj}^2 = 0.5924$. (i) For X_1 : $t_{STAT} = 6.5271$, $p\text{-value} = 0.0000 < 0.05$. Reject H_0 . The assessed value makes a significant contribution and should be included in the model. For X_2 : $t_{STAT} = 0.3617$, $p\text{-value} = 0.7204 > 0.05$. Do not reject H_0 . The age of a house does not make a significant contribution and should not be included in the model. Based on these results, the regression model with only assessed value should be used. (j) For X_1 : $p\text{-value} = 0.0000$. The probability of obtaining a sample that will yield a test statistic farther away than 6.5271 is 0.0000 if the assessed value does not make a significant contribution, holding age constant. For X_2 : $p\text{-value} = 0.7204$. The probability of obtaining a sample that will yield a test statistic farther away than 0.3617 is 0.7204 if the age of a house does not make a significant contribution, holding the effect of the assessed value constant.

(k) $5.9005 \leq \beta_1 \leq 11.3112$. You are 95% confident that the mean taxes will increase by an amount somewhere between \$5.90 and \$11.31 for each additional \$1,000 increase in the assessed value, holding constant the age. In Problem 13.77, you are 95% confident that the mean taxes will increase by an amount somewhere between \$5.91 and \$11.07 for each additional \$thousand increase in assessed value, regardless of the age.

(l) $r_{Y1,2}^2 = 0.6121$. Holding constant the effect of age, 61.21% of the variation in taxes can be explained by variation in the assessed value. $r_{Y2,1}^2 = 0.048$. Holding constant the effect of the assessed value, 0.48% of the variation in taxes can be explained by variation in the age. (m) Based on your answers to (b) through (k), the age of a house does not have an effect on its taxes.

14.82 (a) $\hat{Y} = 163.8291 - 20.0106X_1 - 4.8622X_2$, where $X_1 = \text{ERA}$ and $X_2 = \text{league}$ (American = 0 National = 1). (b) Holding constant the effect of the league, for each additional ERA, the number of wins is estimated to decrease by 20.0106. For a given ERA, a team in the National League is estimated to have 4.8622 fewer wins than a team in the American League. (c) 73.7813 wins. (d) Based on a residual analysis, there is no pattern in the errors. There is no apparent violation of other assumptions. (e) $F_{STAT} = 31.7812 > 3.35$, $p\text{-value} = 0.0000$. Because $p\text{-value} < 0.05$, reject H_0 . There is evidence of a significant relationship between wins and the two independent variables (ERA and league). (f) For X_1 : $t_{STAT} = -7.9253 < -2.0518$, the $p\text{-value} = 0.0000$. Reject H_0 . ERA makes a significant contribution and should be included in the model. For X_2 : $t_{STAT} = -1.9487 > -2.0518$, $p\text{-value} = 0.0618 > 0.05$. Do not reject H_0 . The league does not make a significant contribution and should not be included in the model. Based on these results, the regression model with only the ERA as the independent variable should be used. (g) $-25.1913 \leq \beta_1 \leq -14.83$. (h) $-9.9816 \leq \beta_2 \leq 0.2573$. (i) $r_{adj}^2 = 0.6798$. 67.98% of the variation in wins can be explained by the variation in ERA and league after adjusting for number of independent variables and sample size. (j) $r_{Y1,2}^2 = 0.6994$. Holding constant the effect of league, 69.94% of the variation in number of wins can be explained by the variation in ERA. $r_{Y2,1}^2 = 0.1233$. Holding constant the effect of ERA, 12.33% of the variation in number of wins can be explained by the variation in league. (k) The slope of the number of wins with ERA is the same, regardless of whether the team belongs to the American League or the National League. (l) For X_1X_2 : $t_{STAT} = -0.7514 > -2.0555$ the $p\text{-value}$ is $0.4592 > 0.05$. Do not reject H_0 . There is no evidence that the interaction term makes a contribution to the model. (m) The model with one independent variable (ERA) should be used.

14.84 The r^2 of the multiple regression is very low, at 0.0645. Only 6.45% of the variation in thickness can be explained by the variation of pressure and temperature. The F test statistic for the model including pressure and temperature is 1.621, with $p\text{-value} = 0.2085$. Hence, at a

5% level of significance, there is not enough evidence to conclude that pressure and/or temperature affect thickness. The $p\text{-value}$ of the t test for the significance of pressure is $0.8307 > 0.05$. Hence, there is insufficient evidence to conclude that pressure affects thickness, holding constant the effect of temperature. The $p\text{-value}$ of the t test for the significance of temperature is 0.0820, which is also > 0.05 . There is insufficient evidence to conclude that temperature affects thickness at the 5% level of significance, holding constant the effect of pressure. Hence, neither pressure nor temperature affects thickness individually.

The normal probability plot does not suggest any potential violation of the normality assumption. The residual plots do not indicate potential violation of the equal variance assumption. The temperature residual plot, however, suggests that there might be a nonlinear relationship between temperature and thickness.

The r^2 of the multiple regression model that includes the interaction of pressure and temperature is very low, at 0.0734. Only 7.34% of the variation in thickness can be explained by the variation of pressure, temperature, and the interaction of the two. The F test statistic for the model that includes pressure and temperature and their interaction is 1.214, with a $p\text{-value}$ of 0.3153. Hence, at a 5% level of significance, there is insufficient evidence to conclude that pressure, temperature, and the interaction of the two affect thickness. The $p\text{-value}$ of the t test for the significance of pressure, temperature, and the interaction term are 0.5074, 0.4053, and 0.5111, respectively, which are all greater than 5%. Hence, there is insufficient evidence to conclude that pressure, temperature, or the interaction individually affects thickness, holding constant the effect of the other variables.

The pattern in the normal probability plot and residual plots is similar to that in the regression without the interaction term. Hence the article's suggestion that there is a significant interaction between the pressure and the temperature in the tank cannot be validated.

14.86 $b_0 = 18.2892$ (die temperature), $b_1 = 0.5976$, (die diameter), $b_2 = -13.5108$. The r^2 of the multiple regression is 0.3257 so 32.57% of the variation in unit density can be explained by the variation of die temperature and die diameter. The F test statistic for the combined significance of die temperature and die diameter is 5.0718 with a $p\text{-value}$ of 0.0160. Hence, at a 5% level of significance, there is enough evidence to conclude that die temperature and die diameter affect unit density. The $p\text{-value}$ of the t test for the significance of die temperature is 0.2117, which is greater than 5%. Hence, there is insufficient evidence to conclude that die temperature affects unit density holding constant the effect of die diameter. The $p\text{-value}$ of the t test for the significance of die diameter is 0.0083, which is less than 5%. There is enough evidence to conclude that die diameter affects unit density at the 5% level of significance holding constant the effect of die temperature. After removing die temperature from the model, $b_0 = 107.9267$ (die diameter), $b_1 = -13.5108$. The r^2 of the multiple regression is 0.2724. So 27.24% of the variation in unit density can be explained by the variation of die diameter. The $p\text{-value}$ of the t test for the significance of die diameter is 0.0087, which is less than 5%. There is enough evidence to conclude that die diameter affects unit density at the 5% level of significance. There is some lack of equality in the residuals and some departure from normality. None of the observations have a Cook's $D_i > F_\alpha = 0.8002$ with $df = 2$ and 22. Hence, using the Studentized deleted residuals, hat matrix diagonal elements, and Cook's distance statistic together, there is insufficient evidence for removal of any observation from the model.

CHAPTER 15

15.2 (a) Predicted HOCS is 2.8600, 3.0342, 3.1948, 3.3418, 3.4752, 3.5950, 3.7012, 3.7938, 3.8728, 3.9382, 3.99, 4.0282, 4.0528, 4.0638, 4.0612, 4.045, 4.0152, 3.9718, 3.9148, 3.8442, and 3.76. (c) The

curvilinear relationship suggests that HOCS increases at a decreasing rate. It reaches its maximum value of 4.0638 at GPA = 3.3 and declines after that as GPA continues to increase. (d) An r^2 of 0.07 and an adjusted r^2 of 0.06 tell you that GPA has very low explanatory power in identifying the variation in HOCS. You can tell that the individual HOCS scores are scattered widely around the curvilinear relationship.

15.4 (a) $\hat{Y} = -4.628 + 21.2075X_1 + 4.004X_2$ where X_1 = alcohol % and X_2 = carbohydrates. $F_{STAT} = 2,719.1957$ $p\text{-value} = 0.0000 < 0.05$, so reject H_0 . At the 5% level of significance, the linear terms are significant together. (b) $\hat{Y} = 9.8109 + 15.2538X_1 + 4.466X_2 + 0.477X_1^2 - 0.017X_2^2$, where X_1 = alcohol % and X_2 = carbohydrates. (c) $F_{STAT} = 5.526$ $p\text{-value} = 0.0000 < 0.05$, so reject H_0 . At the 5% level of significance, the model with quadratic terms are significant. $t_{STAT} = 3.3086$, and the $p\text{-value} = 0.0012$. Reject H_0 . There is enough evidence that the quadratic term for alcohol % is significant at the 5% level of significance. $t_{STAT} = -1.1674$, $p\text{-value} = 0.2449$. Do not reject H_0 . There is insufficient evidence that the quadratic term for carbohydrates is significant at the 5% level of significance. Hence, since the quadratic term for alcohol is significant, the model in (b) that includes this term is better. The normal probability plot suggests some left-skewness in the errors. However, because of the large sample size, the validity of the results is not seriously impacted. The residual plots of the alcohol percentage and carbohydrates in the quadratic model do not reveal any remaining nonlinearity. (d) The number of calories in a beer depends quadratically on the alcohol percentage but linearly on the number of carbohydrates. The alcohol percentage and number of carbohydrates explain about 97.454% of the variation in the number of calories in a beer.

15.6 (b) Predicted cost = 710.00 + 607.9773 units - 1.3693 units².
(c) Predicted cost = 710.00 + 607.9773(145) - 1.3693(145)² = \$60,076.79. (d) There appears to be a curvilinear pattern in the residual plot of the units and the units squared. (e) $F_{STAT} = 320.5955 > 4.74$; reject H_0 . (f) $p\text{-value} = 0.0000 < 0.05$, so the model is significant. (g) $t_{STAT} = -5.5790 < -2.3646$; reject H_0 . There is a significant quadratic effect. (h) $p\text{-value} = 0.0008 < 0.05$, so the quadratic term is significant. (i) 98.92% of the variation in yield can be explained by the quadratic model. (j) 98.61%.

15.8 (a) 215.37. (b) For each additional unit of the logarithm of X_1 , the logarithm of Y is estimated to increase by 0.9 unit, holding all other variables constant. For each additional unit of the logarithm of X_2 , the logarithm of Y is estimated to increase by 1.41 units, holding all other variables constant.

15.10 (a) $\hat{Y} = -154.089 + 95.2628\sqrt{X_1} + 27.35\sqrt{X_2}$, where X_1 = alcohol % and X_2 = carbohydrates. (b) The normal probability plot suggests that the errors are right-skewed. However, because of the large sample size, the validity of the results is not seriously impacted. The residual plots of the square-root transformation of alcohol percentage and carbohydrates reveal some remaining nonlinearity. (c) $F_{STAT} = 1,158.2378$. Because the $p\text{-value}$ is 0.0000, reject H_0 at the 5% level of significance. There is evidence of a significant linear relationship between calories and the square root of the percentage of alcohol and the square root of the number of carbohydrates. (d) $r^2 = 0.9396$. So 93.96% of the variation in calories can be explained by the variation in the square root of the percentage of alcohol and the square root of the number of carbohydrates. (e) Adjusted $r^2 = 0.9388$. (f) The model in Problem 15.4 is slightly better because it has a higher r^2 .

15.12 (a) Predicted $\ln(\text{Cost}) = 9.7664 + 0.0080 \text{ Units}$ (b) \$55,471.75.
(c) A quadratic pattern exists, so the model is not adequate.
(d) $t_{STAT} = 7.362 > 2.306$; reject H_0 . (e) 87.14%. 87.14% of the variation in the natural log of the cost can be explained by the number of units.

(f) 85.53%. (g) Choose the model from Problem 15.6. That model has a much higher adjusted r^2 of 98.61%.

15.14 1.25.

$$\mathbf{15.16} R_1^2 = 0.64, VIF_1 = \frac{1}{1 - 0.64} = 2.778, R_2^2 = 0.64,$$

$$VIF_2 = \frac{1}{1 - 0.64} = 2.778. \text{ There is no evidence of collinearity.}$$

15.18 $VIF = 1.0 < 5$. There is no evidence of collinearity.

15.20 $VIF = 1.0105$. There is no evidence of collinearity.

15.22 (a) 35.04. (b) $C_p > 3$. This does not meet the criterion for consideration of a good model.

15.24 Let Y = assessed value, X_1 = size, X_2 = fireplace (0 = no 1 = yes and X_3 = number of bedrooms. X_4 = number of bathrooms. Based on a full regression model involving all of the variables, all the VIF values (2.3355, 1.1873, 1.9885, and 1.4428, respectively) are less than 5. There is no reason to suspect the existence of collinearity. Based on a best-subsets regression and examination of the resulting C_p values, the best models appear to be a model with variables X_1 and X_2 , which has $C_p = 1.1712$, and the regression model with only X_1 . Based on a stepwise regression analysis with all the original variables, only variable X_1 makes a significant contribution to the model at the 0.05 level. Thus, the best model is the model using the size of the house (X_1) as the independent variable. A residual analysis shows no strong patterns. The final model is $\hat{Y} = 276.848 + 508.031X_1$, $r^2 = 0.3273$, $r_{adj}^2 = 0.3033$. Overall significance of the model: $F_{STAT} = 13.326$, $p < 0.001$.

15.30 (a) An analysis of the linear regression model with all of the three possible independent variables reveals that the highest VIF is only 1.06. A stepwise regression model selects only the supplier dummy variable for inclusion in the model. A best-subsets regression produces only one model that has a C_p value less than or equal to $k + 1$ which is the model that includes pressure and the supplier dummy variable. This model is $\hat{Y} = -31.5929 + 0.7879X_2 + 13.1029X_3$. This model has $F = 5.1088$ (2 and 11 degrees of freedom) with a $p\text{-value} = 0.027$. $r^2 = 0.4816$, $r_{adj}^2 = 0.3873$. A residual analysis does not reveal any strong patterns. The errors appear to be normally distributed.

15.32 (a) Best model: $C_p = 2.1558$, predicted fair market value = 260.6791 + 362.8318 land + 0.1109 house size (sq ft) - 1.7543 age. (b) The adjusted r^2 for the best model in 15.32 (a), 15.33 (a), and 15.34 (a) are, respectively, 0.8242, 0.9047, and 0.8481. The model in 15.33 (a) has the highest explanatory power after adjusting for the number of independent variables and sample size.

15.34 (a) Predicted fair market value = 145.1217 + 149.9337 land + 0.0913 house size (sq. ft.). (b) The adjusted r^2 for the best model in 15.32(a), 15.33(a), and 15.34 (a) are, respectively, 0.8242, 0.9047, and 0.8481. The model in 15.33 (a) has the highest explanatory power after adjusting for the number of independent variables and sample size.

15.36 Let Y = fair market value, X_1 = land area, X_2 = interior, X_3 = age, X_4 = number of rooms, X_5 = number of bathrooms, X_6 = garage size, $X_7 = 1$ if Glen Cove and 0 otherwise, and $X_8 = 1$ if Roslyn and 0 otherwise. (a) The VIF s of X_2 , X_3 , and X_7 are greater than 5. Dropping X_2 with the largest VIF , X_3 still has a VIF greater than 5. After dropping X_2 and X_3 , all remaining VIF s are less than 5 so there is no reason to suspect collinearity between any pair of variables. The following is the multiple regression model that has the smallest C_p (4.3211) and the highest adjusted r^2 (0.6815):

$$\begin{aligned} \text{Fair Market Value} = & 49.2379 + 579.0105 \text{ Land} + 109.5767 \text{ Baths} \\ & + 48.2282 \text{ Garage} + 213.2326 \text{ Roslyn} \end{aligned}$$

The individual *t* test for the significance of each independent variable at the 5% level of significance concludes that only property size, baths, and the dummy variable Roslyn are significant given that the others are in the model. The following is the multiple regression result for the model chosen by stepwise regression:

$$\text{Fair Market Value} = 30.3016 + 611.6910 \text{ Land} + 130.7788 \text{ Baths} + 214.2567 \text{ Roslyn}$$

All the variables are significant individually at the 5% level of significance. Combining the stepwise regression and the best-subsets regression results along with the individual *t* test results, the most appropriate multiple regression model for predicting the fair market value is the stepwise regression model. (b) The estimated fair market value in Roslyn is \$214.2567 thousands above Glen Cove or Freeport for two otherwise identical properties.

15.38 In the multiple regression model with catalyst, pH, pressure, temperature, and voltage as independent variables, none of the variables has a *VIF* value of 5 or larger. The best-subsets approach showed that only the model containing X_1, X_2, X_3, X_4 , and X_5 should be considered, where $X_1 = \text{catalyst}$, $X_2 = \text{pH}$, $X_3 = \text{pressure}$, $X_4 = \text{temp}$, and $X_5 = \text{voltage}$. Looking at the *p*-values of the *t* statistics for each slope coefficient of the model that includes X_1 through X_5 reveals that pH level is not significant at the 5% level of significance (*p*-value = 0.2862). The multiple regression model with pH level deleted shows that all coefficients are significant individually at the 5% level of significance. The best linear model is determined to be $\hat{Y} = 3.6833 + 0.1548X_1 - 0.04197X_3 - 0.4036X_4 + 0.4288X_5$. The overall model has $F = 77.0793$ (4 and 45 degrees of freedom), with a *p*-value that is virtually 0. $r^2 = 0.8726$, $r^2_{adj} = 0.8613$. The normal probability plot does not suggest possible violation of the normality assumption. A residual analysis reveals a potential nonlinear relationship in temperature. The *p*-value of the squared term for temperature (0.1273) in the following quadratic transformation of temperature does not support the need for a quadratic transformation at the 5% level of significance. The *p*-value of the interaction term between pressure and temperature (0.0780) indicates that there is not enough evidence of an interaction at the 5% level of significance. The best model is the one that includes catalyst, pressure, temperature, and voltage, which explains 87.26% of the variation in thickness.

CHAPTER 16

16.2 (a) 1988. **(b)** The first four years and the last four years.

16.4 (b), (c), (e)

Drive-Thru

Year	Speed	MA(3)	ES($W = 0.5$)	ES($W = 0.25$)
1998	177.59	177.5900	177.5900	
1999	167.02	171.4967	172.3050	174.9475
2000	169.88	169.2500	171.0925	173.6806
2001	170.85	167.8167	170.9713	172.9730
2002	162.72	163.4967	166.8456	170.4097
2003	156.92	157.3867	161.8828	167.0373
2004	152.52	159.1133	157.2014	163.4080
2005	167.90	161.4400	162.5507	164.5310
2006	163.90	166.3000	163.2254	164.3732
2007	167.10	163.2567	165.1627	165.0549
2008	158.77	166.6967	161.9663	163.4837
2009	174.22	170.753	168.0932	166.1678
2010	179.27	179.2300	173.6816	169.4433
2011	184.20	184.1000	178.9408	173.1325
2012	188.83	183.8854	177.0569	

(d) $W = 0.5: \hat{Y}_{2013} = E_{2012} = 183.8854; W = 0.25: \hat{Y}_{2013} = 177.0569$. **(f)** The exponentially smoothed forecast for 2013 with $W = 0.5$ is higher than that with $W = 0.25$. A smoothing coefficient of $W = 0.25$ smooths out the average time more than $W = 0.50$. The exponential smoothing with $W = 0.5$ assigns more weight to the more recent values and is better for forecasting, while the exponential smoothing with $W = 0.25$, which assigns more weight to more distant values, is better suited for eliminating unwanted cyclical and irregular variations.

16.6 (b), (c), (e)

Performance

Decade	(%)	MA(3)	ES($W = 0.5$)	ES($W = 0.25$)
1830s	2.8		2.8000	2.8000
1840s	12.8	7.4000	7.8000	5.3000
1850s	6.6	10.6333	7.2000	5.6250
1860s	12.5	8.8667	9.8500	7.3438
1870s	7.5	8.6667	8.6750	7.3828
1880s	6.0	6.3333	7.3375	7.0371
1890s	5.5	7.4667	6.4188	6.6528
1900s	10.9	6.2000	8.6594	7.7146
1910s	2.2	8.8000	5.4297	6.3360
1920s	13.3	4.4333	9.3648	8.0770
1930s	-2.2	6.9000	3.5824	5.5077
1940s	9.6	8.5333	6.5912	6.5308
1950s	18.2	12.0333	12.3956	9.4481
1960s	8.3	11.0333	10.3478	9.1611
1970s	6.6	10.5000	8.4739	8.5208
1980s	16.6	13.6000	12.5370	10.5406
1990s	17.6	11.2333	15.0685	12.3055
2000s	-0.5		7.2842	9.1041

(d) $\hat{Y}_{2010} = E_{2000} = 7.2842$ **(e)** $\hat{Y}_{2010} = E_{2000} = 9.1041$. **(f)** The exponentially smoothed forecast for the 2010s with $W = 0.5$ is lower than that with $W = 0.25$. The exponential smoothing with $W = 0.5$ assigns more weight to the more recent values and is better for forecasting, while the exponential smoothing with $W = 0.25$ which assigns more weight to more distant values is better suited for eliminating unwanted cyclical and irregular variations. **(g)** According to the exponential smoothing with $W = 0.25$, there appears to be a general upward trend in the performance of the stocks in the past.

16.8 (b), (c), (e)

Year	Audits	MA(3)	ES($W = 0.5$)	ES($W = 0.25$)
2001	3,305		3,305.00	3,305.00
2002	3,749	3,461.33	3,527.00	3,416.00
2003	3,330	3,821.67	3,428.50	3,394.50
2004	4,386	4,191.67	3,907.25	3,642.38
2005	4,859	4,407.00	4,383.13	3,946.53
2006	4,276	4,186.33	4,329.56	4,028.90
2007	3,424	3,784.67	3,876.78	3,877.67
2008	3,654	3,616.33	3,765.39	3,821.76
2009	3,771	3,625.00	3,768.20	3,809.07
2010	3,450	3,630.00	3,609.10	3,719.30
2011	3,669	3,736.00	3,639.05	3,706.72
2012	4,089		3,864.02	3,802.29

(d) $W = 0.5: \hat{Y}_{2013} = E_{2012} = 3,864.02; W = 0.25: \hat{Y}_{2013} = E_{2012} = 3,802.29$. **(f)** The exponentially smoothed forecast for 2013 with $W = 0.5$ is higher than that with $W = 0.25$. The exponential smoothing with $W = 0.5$ assigns more weight to the more recent values and is better for forecasting, while the exponential smoothing with $W = 0.25$, which assigns more weight to more distant values, is better suited for eliminating unwanted cyclical and irregular variations.

16.10 (a) The Y intercept $b_0 = 4.0$ is the fitted trend value reflecting the real total revenues (in millions of dollars) during the origin, or base year, 1992. **(b)** The slope $b_1 = 1.5$ indicates that the real total revenues are increasing at an estimated rate of \$1.5 million per year. **(c)** Year is 1996, $X = 1996 - 1992 = 4$, $\hat{Y}_5 = 4.0 + 1.5(4) = 10.0$ million dollars. **(d)** Year is 2013, $X = 2013 - 1992 = 21$, $\hat{Y}_{20} = 4.0 + 1.5(21) = 35.5$ million dollars. **(e)** Year is 2016, $X = 2016 - 1992 = 24$, $\hat{Y}_{23} = 4.0 + 1.5(24) = 40$ million dollars.

16.12 (b) Linear trend: $\hat{Y} = 35.9216 + 82.5686X$, where X is relative to 1997. **(c)** Quadratic trend: $\hat{Y} = 53.6925 + 75.4603X + 0.4443X^2$, where X is relative to 1997. **(d)** Exponential trend: $\log_{10}\hat{Y} = 2.1981 + 0.0671X$, where X is relative to 1997. **(e)** Linear trend: $\hat{Y}_{2014} = 35.9216 + 82.5686(17) = 1,439.5882 = 1,440$
 $\hat{Y}_{2015} = 35.9216 + 82.5686(18) = 1,522.1569 = 1,522$

$$\begin{aligned}\text{Quadratic trend: } \hat{Y}_{2014} &= 53.6925 + 75.4603(17) + 0.4443(17)^2 \\ &= 1,464.9118 = 1,465 \\ \hat{Y}_{2015} &= 53.6925 + 75.4603(18) + (0.4443)(18)^2 = 1,555.9216 = 1,556\end{aligned}$$

$$\begin{aligned}\text{Exponential trend: } \hat{Y}_{2014} &= 10^{2.1981+0.0671(17)} = 2,180.1029 = 2,180 \\ \hat{Y}_{2015} &= 10^{2.1981+0.0671(18)} = 2,544.2451 = 2,544.\end{aligned}$$

(f) The linear and quadratic trend model fit the data better than the exponential trend model and, hence, either forecast should be used.

16.14 (b) $\hat{Y} = 315.7662 + 65.2175X$ where X = years relative to 1978. **(c)** $X = 2013 - 1978 = 35$, $\hat{Y} = 315.7662 + 65.2175(35) = \$2,598.3770$ billion $X = 2014 - 1978 = 36$, $\hat{Y} = 3,105.7662 + 65.2175(36) = \$2,663.5944$ billion. **(d)** There is an upward trend in federal receipts between 1978 and 2012. The trend appears to be linear.

16.16 (b) Linear trend: $\hat{Y} = -95 + 245.49098X$, where X is relative to 2002. **(c)** Quadratic trend: $\hat{Y} = 930.2098 - 437.9823X + 68.3473X^2$, where X is relative to 2002. **(d)** Exponential trend: $\log_{10}\hat{Y} = 2.5659 + 0.0747X$, where X is relative to 2002. **(e)** Linear trend:
 $\hat{Y}_{2013} = -95.0000 + 245.4909(11) = 2,605.4$ million KWh
 $\hat{Y}_{2014} = -95.0000 + 245.4909(12) = 2,850.89$ million KWh
Quadratic trend: $\hat{Y}_{2013} = 930.2098 - 437.9823(11) + 68.3473(11)^2 = 4,382.43$ million KWh
 $\hat{Y}_{2014} = 930.2098 - 437.9823(12) + 68.3473(12)^2 = 5,516.44$ millions of KWh
Exponential trend: $\hat{Y}_{2013} = 10^{2.5659+0.0747(11)} = 2,438.44$ million KWh
 $\hat{Y}_{2014} = 10^{2.5659+0.0747(12)} = 2,895.82$ million KWh.

16.18 (b) Linear trend: $\hat{Y} = 2.0500 + 0.1321X$, where X is relative to 2000. **(c)** Quadratic trend: $\hat{Y} = 2.1471 + 0.0835X - 0.0037X^2$, where X is relative to 2000. **(d)** Exponential trend: $\log_{10}\hat{Y} = 0.3274 + 0.0198X$, where X is relative to 2000.

(e)

1st Difference	2nd Difference	%Difference
0.30		15.08
0.09	-0.21	3.93
0.20	0.11	8.40
-0.09	-0.29	-3.49
0.14	0.23	5.62
0.20	0.06	7.60
0.09	-0.11	3.18
0.21	0.12	7.19
0.13	-0.08	4.15
0.01	-0.12	0.31
0.05	0.04	1.53
0.06	0.01	1.81
0.87	0.81	25.74

The linear and exponential models should be investigated.

(f) The forecasts using the two models are:

Linear trend: $\hat{Y}_{2014} = 2.0500 + 0.1321(14) = \3.8992 millions

Exponential trend: $\hat{Y}_{2014} = 10^{0.3274+0.0198(14)} = \4.0184 millions

16.20 (b) There has been an upward trend in the CPI in the United States over the 48-year period. The rate of increase became faster in the late 70s but tapered off in the early 80s.

(c) Linear trend: $\hat{Y} = 16.2540 + 4.4919X$. **(d)** Quadratic trend:

$\hat{Y} = 19.7334 + 4.0381X + 0.0097X^2$. **(e)** Exponential trend:

$\log_{10}\hat{Y} = 1.5612 + 0.0192X$. **(f)** None of the trend models appear appropriate according to the first-, second- and percentage-difference but the second-difference has the smallest variation in general. So a quadratic trend model is slightly preferred over the others. **(g)** Quadratic trend:

For 2013: $\hat{Y}_{2013} = 19.7334 + 4.0381(48) + 0.0097(48)^2 = 235.8101$

For 2014: $\hat{Y}_{2014} = 19.7334 + 4.0381(49) + 0.0097(49)^2 = 240.7848$

16.22 (a) For Time Series I, the graph of Y versus X appears to be more linear than the graph of $\log Y$ versus X , so a linear model appears to be more appropriate. For Time Series II, the graph of $\log Y$ versus X appears to be more linear than the graph of Y versus X , so an exponential model appears to be more appropriate.

(b) Time Series I: $\hat{Y} = 100.0731 + 14.9776X$, where X = years relative to 2001

Time Series II: $\hat{Y} = 10^{1.9982+0.0609X}$, where X = years relative to 2001.

(c) $X = 12$ for year 2012 in all models. Forecasts for the year 2012:

Time Series I: $\hat{Y} = 100.0731 + 14.9776(12) = 279.8045$

Time Series II: $\hat{Y} = 10^{1.9982+0.0609(12)} = 535.6886$.

16.24 $t_{STAT} = 2.40 > 2.2281$; reject H_0 .

16.26 (a) $t_{STAT} = 1.60 < 2.2281$; do not reject H_0 .

16.28 (a)

Standard			
Coefficients	Error	t Stat	P-value
Intercept	73.0383	79.1710	0.9225
YLag1	2.2358	1.5648	1.4288
YLag2	-2.8050	2.9255	-0.9588
YLag3	1.6443	1.5045	1.0929

Since the p -value = 0.3001 > 0.05 level of significance, the third-order term can be dropped.

(b)

Standard			
Coefficients	Error	t Stat	P-value
Intercept	70.1584	60.1327	1.1667
YLag1	0.9157	0.6571	1.3935
YLag2	0.1243	0.6541	0.1900

Since the p -value = 0.8525 and is greater than 0.05, the second-order term can be dropped.

(c)

Standard			
Coefficients	Error	t Stat	p-value
Intercept	53.9415	31.1693	1.7306
YLag1	1.0482	0.0418	25.0917

Since the p -value = 0.0000, the first-order term cannot be dropped.

(d) The most appropriate model for forecasting is the first-order autoregressive model:

$$\begin{aligned}\hat{Y}_{2014} &= 53.9415 + 1.0482Y_{2013} = 1,596 \text{ stores.} \\ \hat{Y}_{2015} &= 53.9415 + 1.0482\hat{Y}_{2014} = 1,727 \text{ stores.}\end{aligned}$$

16.30 (a)

Standard			
Coefficients	Error	t Stat	P-value
Intercept	-0.1927	0.7859	-0.2452
YLag1	0.4966	1.0298	0.4822
YLag2	0.3137	1.1551	0.2716
YLag3	0.3507	0.8621	0.4068
			0.7938
			0.6963

Since the p -value = 0.6963 > 0.05 level of significance, the third-order term can be dropped.

(b)

Standard			
Coefficients	Error	t Stat	P-value
Intercept	-0.2268	0.6014	-0.3771
YLag1	0.7751	0.8212	0.9438
YLag2	0.3759	0.7459	0.5039
			0.3699
			0.6264

Since the p -value = 0.6264 > 0.05 level of significance, the second-order term can be dropped.

(c)

Standard			
Coefficients	Error	t Stat	P-value
Intercept	-0.0479	0.4358	-0.1100
YLag1	1.0791	0.1535	7.0278
			0.9144
			0.0000

Since the p -value is 0.0000, the first-order term cannot be dropped.

(d) The most appropriate model for forecasting is the first-order autoregressive model:

$$\hat{Y}_{2014} = -0.0479 + 1.0791Y_{2013} = \$4.5380 \text{ million.}$$

16.32 (a) 2.121. **(b)** 1.50.

16.34 (a) The residuals in the linear and exponential trend model show strings of consecutive positive and negative values.

(b), (c)

	Linear	Quadratic	Exponential	AR1
SSE	6,275,572.218	2,267,548.717	5,613,646	760,222.4864
Syx	835.0364	532.3942	789.7711	308.2658
MAD	580.8397	367.8580	364.4034	222.0914

(d) The residuals in the three trend models show strings of consecutive positive and negative values. The autoregressive model performs well for the historical data and has a fairly random pattern of residuals. It has the smallest values in MAD and S_{YX} . Based on the principle of parsimony, the autoregressive model would be the best model for forecasting.

16.36 (b), (c)

	Linear	Quadratic	Exponential	AR1
SSE	43,144.3137	41,614.2394	547,380.1537	53,441.7002
Syx	53.6310	54.5201	191.0288	61.7840
MAD	40.1915	40.6096	133.0124	37.0922

(d) The residuals in the three trend models show strings of consecutive positive and negative values. The autoregressive model performs well for the historical data and has a fairly random pattern of residuals. The autoregressive model also has the smallest values in MAD and S_{YX} . Based on the principle of parsimony, the autoregressive model would be the best model for forecasting.

16.38 (b), (c)

	Linear	Quadratic	Exponential	AR1
SSE	0.3977	0.3571	0.3538	0.6299
Syx	0.1821	0.1802	0.1717	0.2393
MAD	0.1193	0.1182	0.1165	0.1490

(d) The residuals in the linear, quadratic, and exponential trend models show strings of consecutive positive and negative values. The autoregressive model performs well for the historical data and has a fairly random pattern of residuals. The exponential trend model, however, has the smallest values in MAD and S_{YX} . The autoregressive model would be the best model for forecasting due to its fairly random pattern of residuals even though it has slightly larger MAD and S_{YX} than the exponential model.

16.40 (a) $\log \hat{\beta}_0 = 2$, $\hat{\beta}_0 = 10^2 = 100$. This is the fitted value for January 2009 prior to adjustment with the January multiplier.

(b) $\log \hat{\beta}_1 = 0.01$, $\hat{\beta}_1 = 10^{0.01} = 1.0233$. The estimated monthly compound growth rate is 2.33%. **(c)** $\log \hat{\beta}_2 = 0.1$, $\hat{\beta}_2 = 10^{0.1} = 1.2589$. The January values in the time series are estimated to have a mean 25.89% higher than the December values.

16.42 (a) $\log \hat{\beta}_0 = 3.0$, $\hat{\beta}_0 = 10^{3.0} = 1,000$. This is the fitted value for the first quarter of 2009 prior to adjustment by the quarterly multiplier.

(b) $\log \hat{\beta}_1 = 0.1$, $\hat{\beta}_1 = 10^{0.1} = 1.2589$. The estimated quarterly compound growth rate is $(\hat{\beta}_1 - 1)100\% = 25.89\%$.

(c) $\log \hat{\beta}_3 = 0.2$, $\hat{\beta}_3 = 10^{0.2} = 1.5849$.

16.44 (a) The retail industry is heavily subject to seasonal variation due to the holiday seasons and so are the revenues for Toys R Us.

(b) There is obvious seasonal effect in the time series.

(c) $\log_{10} \hat{Y} = 3.6291 + 0.0025X - 0.3670Q_1 - 0.3676Q_2 - 0.3438Q_3$.

(d) $\log_{10} \hat{\beta}_1 = 0.0025$. $\hat{\beta}_1 = 10^{0.0025} = 1.0058$. The estimated quarterly compound growth rate is $(\hat{\beta}_1 - 1)100\% = 0.58\%$.

(e) $\log_{10} \hat{\beta}_2 = -0.3670$. $\hat{\beta}_2 = 10^{-0.3670} = 0.4296$.

$(\hat{\beta}_2 - 1)100\% = -57.04\%$. The 1st quarter values in the time series are estimated to have a mean 57.04% below the 4th quarter values.

$\log_{10} \hat{\beta}_3 = -0.3676$. $\hat{\beta}_3 = 10^{-0.3676} = 0.4289$. $(\hat{\beta}_3 - 1)100\% = -57.11\%$.

The 2nd quarter values in the time series are estimated to have a mean 57.11% below the 4th quarter values.

$\log_{10} \hat{\beta}_4 = -0.3438$. $\hat{\beta} = 10^{-0.3438} = 0.4531$. $(\hat{\beta}_4 - 1)100\% = -54.69\%$.

The 3rd quarter values in the time series are estimated to have a mean 54.69% below the 4th quarter values. **(f)** Forecasts for the last three quarters of 2013 and all of 2014 are 2,723.336, 2,893.926, 6,423.647, 2,775.405, 2,787.204, 2,961.794, and 6,574.295 millions

16.46 (b)

	Standard			
	Coefficients	Error	t Stat	P-value
Intercept	-0.0916	0.0070	-13.1771	0.0000
Coded Month	-0.0037	0.0001	-64.4666	0.0000
M1	0.1245	0.0086	14.5072	0.0000
M2	0.0905	0.0086	10.5454	0.0000
M3	0.0806	0.0086	9.3975	0.0000
M4	0.0668	0.0086	7.7810	0.0000
M5	0.0793	0.0086	9.2457	0.0000
M6	0.1001	0.0086	11.6725	0.0000
M7	0.1321	0.0086	15.3949	0.0000
M8	0.1238	0.0086	14.4279	0.0000
M9	0.0577	0.0088	6.5372	0.0000
M10	0.0356	0.0088	4.0317	0.0001
M11	0.0103	0.0088	1.1646	0.2472

(c) $\hat{Y}_{103} = 0.4451$.

(d) Forecasts for the last four months of 2012 are 0.3790, 0.3571, 0.3340, and 0.3234.

(e) $\log_{10}\hat{\beta}_1 = -0.0037$; $\hat{\beta}_1 = 10^{-0.0037} = 0.9915$. The estimated monthly compound growth rate is $(\hat{\beta}_1 - 1) 100\% = -0.8542\%$.(f) $\log_{10}\hat{\beta}_8 = 0.1321$; $\hat{\beta}_8 = 10^{0.1321} = 1.3554$.(g) $(\hat{\beta}_8 - 1) 100\% = 35.5382\%$. The July values in the time series are estimated to have a mean 35.5382% above the December values.**16.48 (b)**

	Standard			
	Coefficients	Error	t Stat	P-value
Intercept	0.7884	0.0362	21.8058	0.0000
Coded Quarter	0.0213	0.0013	16.5210	0.0000
Q1	0.0545	0.0378	1.4406	0.1597
Q2	0.0044	0.0377	0.1165	0.9080
Q3	-0.0080	0.0377	0.2117	0.8338

(c) $\log_{10}\hat{\beta}_1 = 0.0213$; $\hat{\beta}_1 = 10^{0.0213} = 1.05226$; $(\hat{\beta}_1 - 1) 100\% = 5.0226\%$. The estimated quarterly compound mean growth rate in the price of silver is 5.0226%, after adjusting for the seasonal component.(d) $\log_{10}\hat{\beta}_2 = 0.0545$; $\hat{\beta}_2 = 10^{0.0545} = 1.1337$; $(\hat{\beta}_2 - 1) 100\% = 13.37\%$. The first-quarter values in the time series are estimated to have a mean 13.37% above the fourth-quarter values.(e) Last quarter, 2012: $\hat{Y}_{35} = \$34.1414$.

(f) 2013: 40.6502, 38.0402, 40.2810, 41.5349.

16.60 (b) Linear trend: $\hat{Y} = 174,015.2828 + 2,436.3739X$, where X is relative to 1984.(c) 2013: $\hat{Y}_{2013} = 174,015.2828 + 2,436.3739(29) = 244,670.1256$ thousands2014: $\hat{Y}_{2014} = 174,015.2828 + 2,436.3739(30) = 247,106.4995$ thousands.(d) (b) Linear trend: $\hat{Y} = 115,574.3034 + 1,553.1828X$, where X is relative to 1984.(c) 2012: $\hat{Y}_{2013} = 115,574.3034 + 1,553.1828(29) = 160,616.6034$ thousands.2013: $\hat{Y}_{2014} = 115,574.3034 + 1,5853.1828(30) = 162,169.7862$ thousands.**16.62** Linear trend: $\hat{Y} = -2.5854 + 0.7197X$, where X is relative to 1975.

	Standard			
	Coefficients	Error	t Stat	P-value
Intercept	-2.5854	0.6244	-4.1404	0.0002
Coded Yr	0.7197	0.0290	24.7863	0.0000

(c) Quadratic trend: $\hat{Y} = 1.24247 + 0.0847X + 0.0172X^2$, where X is relative to 1975.

	Standard			
	Coefficients	Error	t Stat	P-value
Intercept	1.2247	0.2405	5.0932	0.0000
Coded Yr	0.0847	0.0301	2.8160	0.0079
Coded Yr Sq	0.0172	0.0008	21.8464	0.0000

Exponential trend: $\log_{10}\hat{Y} = 0.1741 + 0.0374X$, where X is relative to 1975.

	Standard			
	Coefficients	Error	t Stat	P-value
Intercept	0.1741	0.0216	8.0477	0.0000
Coded Yr	0.0374	0.0010	37.1890	0.0000
AR(3): $\hat{Y}_i = 0.4887 + 1.1725Y_{i-1} - 0.8248Y_{i-2} + 0.7121Y_{i-3}$				
Intercept	0.4887	0.1561	3.1301	0.0038
YLag1	1.1725	0.1543	7.5975	0.0000
YLag2	-0.8248	0.2592	-3.1817	0.0033
YLag3	0.7121	0.1833	3.8855	0.0005

Test of A_3 : $p\text{-value} = 0.0005 < 0.05$. Reject H_0 that $A_3 = 0$. Third-order term cannot be deleted. A third-order autoregressive model is appropriate.

	Linear	Quadratic	Exponential	AR3
SSE	138.6905	9.4759	206.5249	7.7632
Syx	1.9628	0.5203	2.3952	0.5004
MAD	1.6711	0.3576	1.3060	0.3682

(h) The residuals in the first three models show strings of consecutive positive and negative values. The autoregressive model performs well for the historical data and has a fairly random pattern of residuals. It also has the smallest values in the standard error of the estimate, MAD and SSE. Based on the principle of parsimony, the autoregressive model would probably be the best model for forecasting.

(i) $\hat{Y}_{2013} = 0.4887 + 1.1725Y_{2012} - 0.8248Y_{2011} + 0.7121Y_{2010} = \27.7445 billions.**CHAPTER 17****17.2 (b)** Gauges take up too much visual space that would be needed for the 20 funds. (c) The one-year return and the three-year return is much higher for the long-term funds than for the short-term funds.**17.4 (c)** The bullet graph enables you to see the names of the individual teams and which teams are inexpensive, typical, and expensive whereas the stem-and-leaf display only shows the distribution of the costs. (d) The bullet graphs shows the costs for each team and which teams fall into the inexpensive, typical, and expensive categories.

17.6 (b) The rates of return of the three indices vary a great deal from year to year, but the pattern for the three indices are similar except for the NASDAQ in 2009 which had a much higher return in that year than the DJIA or the S & P500. **(c)** Unlike the three stock indices which had similar patterns between 2006–2012, the returns of the three metals differed greatly from year to year.

17.8 (b) Wendy's consistently has the fastest service time. The service time at McDonald's and Burger are similar over the years. The service time at Chick-Fil-A was slower in the earlier and later years than the other fast food chains.

17.10 (b) The values of the teams varied from \$315 million for the Charlotte Hornets to \$1,100 million for the New York Knicks. The change in values was not consistent across the teams. The two most valuable teams, the Los Angeles Lakers, and the New York Knicks had very different increases in value (11% and 41% respectively).

17.12 (c) Almost all the countries that had lower GDP had lower Internet use except for the Republic of Korea. The pattern of mobile cellular subscriptions does not seem to depend on the GDP of the country.

17.14 (a)

Type	Star Rating						
	One	Two	Three	Four	Five	Grand Total	
<input checked="" type="checkbox"/> Growth	16	43	74	76	18	227	
Large	5	21	37	31	9	103	
Mid-Cap	4	13	20	28	7	72	
Small	7	9	17	17	2	52	
<input checked="" type="checkbox"/> Value	7	19	36	22	5	89	
Large	5	9	21	13	2	50	
Mid-Cap	0	5	9	4	1	19	
Small	2	5	6	5	2	20	
Grand Total	23	62	110	98	23	316	

(b) There are 37 funds.

17.16 (a)

Type	Star Rating						
	One	Two	Three	Four	Five	Grand Total	
<input checked="" type="checkbox"/> Growth	16	43	74	76	18	227	
Average	6	22	28	15	3	74	
High	5	3	1	1	0	10	
Low	5	18	45	60	15	143	
<input checked="" type="checkbox"/> Value	7	19	36	22	5	89	
Average	3	6	7	0	1	17	
High	2	1	0	0	0	3	
Low	2	12	29	22	4	69	
Grand Total	23	62	110	98	23	316	

(b) There is only one fund.

17.18 The five-year returns for the four funds that are small market cap funds that have a rating of five stars are 5.29, 6.97, 10.75, and 11.35.

17.20 The highest five-year return of 12.33 is for a large cap growth fund.

17.22 (b) The r^2 for the classification tree model is 0.434. The first split is for the 8 customers who called 50 or more times. Among customers who called fewer than 50 times, those who called at least seven times and visited two or more times are more likely to churn.

17.24 Because half the data will be used for a validation sample, the results will differ depending on which values are in the training sample and which are in the validation sample.

17.26 The r^2 for the regression tree model is 0.373. The first split is based on a plate gap of 1.8. For those bags with a plate gap less than 1.8, the mean tear is 0.3107. For those bags with a plate gap at least 1.8, the mean tear is 1.98. For those bags with a plate gap less than 0.0, the mean tear is 0.06. For those bags with a plate gap less than 1.8 but greater than 0, the mean tear is 0.45. Thus, you would recommend that a plate gap of less than 0 be used to minimize tears in the bag.

17.28 The r^2 for the regression tree model is 0.789. The first split is based on 831 square feet. Moves of at least 831 sq. ft. have a mean moving time of 51.1875 hours. Moves of less than 831 square feet have a mean moving time of 22.6071 hours. Among moves of less than 831 sq. ft., moves of less than 486 sq. ft., have a mean moving time of 15.7955 hours. Moves of less than 344 sq. ft. have a mean moving time of 12.75 hours. Moves of between 344 and 486 sq. ft. have a mean moving time of 18.3333 hours. Moves of between 486 and 830 sq. ft. have a mean moving time of 27.0147 hours. Moves between 486 and 599 sq. ft. have a mean moving time of 24.825 hours. Moves between 600 and 830 have a mean moving time of 30.1429 hours. Moves between 557 and 599 sq. ft. have a mean moving time of 24.05 hours. Moves between 486 and 557 sq. ft. have a mean moving time of 25.6 hours.

17.30 Because some of the data will be used for a validation sample, the results will differ depending on which values are in the training sample and which are in the validation sample.

17.32 Because some of the data will be used for a validation sample, the results will differ depending on which values are in the training sample and which are in the validation sample.

17.34 Because some of the data will be used for a validation sample, the results will differ depending on which values are in the training sample and which are in the validation sample.

17.36 Because some of the data will be used for a validation sample, the results will differ depending on which values are in the training sample and which are in the validation sample.

17.38 (b) The first two cereals to cluster are Wheaties and Nature's Path Organic Multigrain Flakes followed by Post Shredded Wheat Vanilla Almond and Kellogg's Mini Wheats. At the two cluster level, one cluster contains Post Shredded Wheat Vanilla Almond and Kellogg's Mini Wheats and the other cluster contains the other five cereals.

17.40 (b) The first two countries to cluster are Egypt and Jordan followed by Lithuania and Poland. At the two cluster level one cluster is France, Germany, Spain, United Kingdom, Japan, Israel, and the United States and the other cluster contains the remaining countries. The first cluster appears to contain the western European countries and the United States and Israel.

17.42 (b) Since the stress statistic is 0.0973 in three dimensions, 0.1308 in two dimensions, and 0.3147 in one dimension, it is reasonable to try to interpret a two-dimensional mapping of the cereals. Looking at a 45° rotation, one dimension separates Post Shredded Wheat Vanilla Almond and Kellogg's Mini Wheats based on their higher calorie and sugar content. A second dimension does not seem to be interpretable. In addition, All Bran, which has lower calories and higher sugar is separated from the other cereals.

17.44 (b) Since the stress statistic is 0.2773 in four dimensions, 0.3151 in three dimensions, 0.3787 in two dimensions, and 0.5323 in one dimension, it is reasonable to try to first try interpret a two-dimensional

mapping of the countries. There does not seem to be a clear interpretation of the dimensions along the lines of GDP and social media usage. Pakistan seems very separated from the other countries with Indonesia on the other side of the graph. Russia and Lithuania are close as are Mexico and Spain, and Israel and Egypt.

17.52 (b) The sparklines show differences in the values of the U. S. dollar in terms of the Canadian dollar, the English pound, and the Euro over the time period 2002–2012. The value of the U. S. dollar in terms of the Canadian dollar declined drastically from 2002 to 2007, but has remained steady since 2009. The value of the U. S. dollar in terms of the English pound has remained relatively steady between 2002 and 2012. The value of the U. S. dollar in terms of the Euro declined between 2002 and 2007 (with an increase in 2005) but has remained steady since 2008.

17.54 Because half the data will be used for a validation sample, the results will differ depending on which values are in the training sample and which are in the validation sample.

17.56 Because half the data will be used for a validation sample, the results will differ depending on which values are in the training sample and which are in the validation sample.

17.58 (c) The first two foods to cluster are Cantonese and American, followed by French and Mandarin, followed by Spanish and Greek. At the two cluster level, the first cluster includes Japanese, French, Mandarin, Szechuan, and Mexican. The second cluster includes Cantonese, American, Spanish, Greek, and Italian. Since the stress statistic is 0.0468 in four dimensions, 0.1164 in three dimensions, 0.2339 in two dimensions, and 0.4079 in one dimension, it is reasonable to try to first try interpret a two-dimensional mapping of the foods. There does not seem to be a clear interpretation of the dimensions along the lines of the three scales. The two spicy foods, Mexican and Szechuan are close to each other as are French and Greek, and Japanese and American. Italian is separated by itself as is Spanish.

This page intentionally left blank

Index

A

α (level of significance), 312
A priori probability, 153
Addition rule, 158
Adjusted r , 550
Akaike information criterion, 685
Algebra, rules for, 716
Alternative hypothesis, 309
Among-block variation, 411
Among-group variation, 397, 411
Analysis of means (ANOM), 404
Analysis of proportions (ANOP), 460
Analysis of variance (ANOVA),
 Kruskal-Wallis rank test for differences in c medians, 473–476
 assumptions of, 476
 One-way,
 assumptions, 405
 F test for differences in more than two means, 398
 F test statistic, 398
 Levene's test for homogeneity of variance, 405–406
 summary table, 399
 Tukey-Kramer procedure, 402–404
 Randomized block,
 Assumptions, 415
 Testing for factor and block effects, 410–415
 F test for factor effect, 412
 F test for block effect, 413
 Two-way,
 cell means plot, 426
 factorial design, 418
 interpreting interaction effects, 426–427
 multiple comparisons, 424–425
 summary table, 422
 testing for factor and interaction effects, 421–422
Analysis ToolPak,
 Checking for presence, 741
 Frequency distribution, 88,
 Histogram, 92
 Descriptive statistics, 146
 Exponential smoothing, 669
 F test for ratio of two variances, 390
 Multiple regression, 589
 One-way ANOVA, 440
 paired t test, 388–389
 pooled-variance t test, 386–387
 randomized block design, 443
 random sampling, 34
 sampling distributions, 270
 separate-variance t test, 387–388
 simple linear regression, 539
 two-way ANOVA, 442–443
Residual analysis, 540
Sampling distributions, 270

Analyze, 2, 14

ANOVA. *See* Analysis of variance (ANOVA)
Area of opportunity, 202
Arithmetic mean. *See* Mean
Arithmetic operations, rules for, 716
Assumptions
 analysis of variance (ANOVA), 405
 of the confidence interval estimate for the mean (σ unknown), 279
 of the confidence interval estimate for the proportion, 287
 of the F test for the ratio of two variances, 375
 of the paired t test, 362–363,
 of Kruskal-Wallis test, 476
 of regression, 507
 for 2×2 table, 453
 for $2 \times c$ table, 466
 for $r \times c$ table, 466
 for the t distribution, 280
 t test for the mean (σ unknown), 324–325
 in testing for the difference between two means, 351
 of the Wilcoxon rank sum test, 467
 of the Z test for a proportion, 332
Autocorrelation, 511
Autoregressive modeling, 647
 steps involved in, on annual time-series data, 651–652
 for trend fitting and forecasting, 647–653
Average linkage, 692

B

Bar chart, 51–52
Bayes' theorem, 169
Best-subsets approach in model building, 612–613
 β Risk, 312
Bias
 nonresponse, 25
 selection, 25
Big data, 4
Binomial distribution, 195–201
 mean of, 200
 properties of, 195
 shape of, 199
 standard deviation of, 200
Binomial probabilities
 calculating, 197–199
Boxplots, 124–125
Brynne packaging, 536–537
Bullet graph, 677
Business analytics, 4, 675

C

CardioGood Fitness, 30–31, 83, 144, 182, 246, 303, 384, 438, 484
Categorical data
 chi-square test for the difference between two proportions, 448–453
 chi-square test of independence, 461–466
 chi-square test for c proportions, 455–458

- organizing, 37–40, 68–69
 visualizing, 51–55
 Z test for the difference between two proportions, 371–372
 Categorical variables, 14
 Causal forecasting methods, 630
 Cell means plot, 426
 Cell, 6
 Central limit theorem, 257
 Central tendency, 102–107
 Certain event, 152
 Challenges in organizing and visualizing variables,
 Obscuring data, 70–71
 Creating false impressions, 71
 Chartjunk, 72–74
 Charts.
 bar, 51–52
 Pareto, 53–55
 pie, 52–53
 side-by-side bar, 55
 Chebyshev Rule, 130
 Chi-square (χ^2) distribution, 450,
 Chi-square (χ^2) test for differences
 between c proportions, 455–458
 between two proportions, 448–454
 Chi-square (χ^2) test for the variance or standard deviation, 478
 Chi-square (χ^2) test of independence, 461–466
 Chi-square (χ^2) table, 748
 Choice Is Yours Followup, 83, 182
 Class boundaries, 43
 Class intervals, 43
 Class midpoint, 44
 Class interval width, 43
 Classes, 43
 And Excel bins, 45
 Classification trees, 683–684
 Clear Mountain State Surveys, 31, 84, 144, 182, 246, 303, 385,
 438, 484
 Cluster analysis, 691
 Cluster sample, 23
 Coefficient of correlation, 134, 519
 inferences about, 519–520
 Coefficient of determination, 503–504
 Coefficient of multiple determination, 549–550, 602
 Coefficient of partial determination, 562
 Coefficient of variation, 112
 Collectively exhaustive events, 157
 Collect, 2, 14
 Collinearity of independent variables, 608–609
 Combinations, 176, 196
 Complement, 156
 Complete linkage, 692
 Completely randomized design, 395. *See also* One-way analysis of variance
 Computing conventions used in this book, 7
 Cook's distance statistic D_i , 579–580
 Conditional probability, 161–162
 Confidence coefficient, 313
 Confidence interval estimation, 273
 connection between hypothesis testing and, 319–320
 for the difference between the means of two independent groups, 353
 for the difference between the proportions of two independent groups, 371–372
 for the mean difference, 365
 ethical issues and, 295–296
 for the mean (σ known), 273–276
 for the mean (σ unknown), 279–285
 for the mean response, 523–524
 for the proportion, 287–289
 of the slope, 518, 556
 Contingency tables, 39, 155, 448
 Continuous probability distributions, 270
 Continuous variables, 15
 Control chart factors,
 tables, 757
 Convenience sampling, 21
 Counting rules, 174–176
 Correlation coefficient. *See* Coefficient of correlation
 Covariance, 132
 of a probability distribution, 190
 Coverage error, 25
 Craybill Instrumentation Company case, 624
 Critical range, 403, 416, 425
 Critical value approach, 314–317
 Critical values, 311
 of test statistic, 310–311
 Cross-product term, 566
 Cross validation, 616
 Cumulative percentage distribution, 47–48
 Cumulative percentage polygons, 61–62
 Cumulative standardized normal distribution, 223,
 tables, 744–745
 Cyclical effect, 666
- D**
- Dashboards, 676
 Data, 3
 sources of, 18
 Data cleaning, 20
 Data collection, 18–21
 Data formatting, 19
 Data mining, 682
 Data discovery, 678
 DCOVA, 2
 Decision trees, 163
 Define, 2, 14
 Degrees of freedom, 281
 Dependent variable, 492
 Descriptive analytics, 675
 Descriptive statistics, 4
 Deviance statistic, 576
 Digital Case, 31–32, 83, 144, 182, 214, 246, 269, 302, 342, 384,
 437–438, 483, 536, 588, 624, 668
 Directional test, 328
 Discrete probability distributions
 binomial distribution, 197
 covariance, 190
 hypergeometric distribution, 206
 Poisson distribution, 203
 Discrete variables, 15
 expected value of, 187
 probability distribution for, 186
 variance and standard deviation of, 187–188
 Dispersion, 107
 Downloading files for this book, 729–736
 Drill-down, 678–679

Dummy variables, 564–566
 Durbin-Watson statistic, 512–514
 tables, 756

E

Electronic formats and encoding, 19
 Empirical probability, 153
 Empirical rule, 129
 Ethical issues
 confidence interval estimation and, 295
 in hypothesis testing, 337
 in multiple regression, 618
 in numerical descriptive measures, 137–138
 for probability, 177
 for surveys, 26
 Euclidean distance, 692, 694
 Events, 153
 Expected frequency, 449
 Expected value, 186
 of discrete variable, 187
 of sum of two variables, 191
 Explained variation or regression sum of squares (SSR), 502–503
 Explanatory variables, 493
 Exponential distribution, 240
 Mean of, 240
 Standard deviation of, 240
 Exponential growth
 with monthly data forecasting equation, 660
 with quarterly data forecasting equation, 661
 Exponential smoothing, 634–636
 Exponential trend model, 641–642
 Extrapolation, predictions in regression analysis and, 497

F

Factor, 395
 Factorial design. *See* Two-way analysis of variance
 F distribution, 398
 tables, 749–752
 Finite population correction facvtor, 207, 265
 First-order autoregressive model, 648
 First quartile, 120
 Five-number summary, 123
 Fixed effects models, 431
 Forecasting,
 autoregressive modeling for, 647–653
 choosing appropriate model for, 655–657
 least-squares trend fitting and, 637–645
 seasonal data, 658–663
 Frame, 21
 Frequency distribution, 43–44
 Friedman rank test, 478
 F test for the ratio of two variances, 374–377
 F test for the block effect, 413
 F test for the factor effect, 412
 F test for factor A effect, 421
 F test for factor B effect, 421
 F test for interaction effect, 422
 F test for the slope, 517–518
 F test in one-way ANOVA, 398

G

Gauges, 677
 Gaussian distribution, 270

General addition rule, 158–159
 General multiplication rule, 168
 Geometric mean, 106
 Geometric mean rate of return, 107
 Grand mean, 396
 Greek alphabet, 721
 Groups, 395
 Guidelines for developing visualizations, 74

H

Hat matrix diagonal elements h_i , 579
 Hierarchical clustering, 691
 Histograms, 59–60
 Homogeneity of variance, 405
 Levene's test for, 405–406
 Homoscedasticity, 507
 Hyperbolic tangent function, 689
 Hypergeometric distribution, 206
 mean of, 207
 standard deviation of, 207
 Hypergeometric probabilities
 calculating, 207–208
 Hypothesis. *See also* One-sample tests of hypothesis
 alternative, 309
 null, 309

I

Impossible event, 152
 Independence, 165
 of errors, 507
 χ^2 test of, 461–466
 Independent events, multiplication rule for, 168
 Independent variable, 492
 Index numbers, 664
 Inferential statistics, 4
 Interaction, 419, 566–567
 Interaction terms, 566
 Interpolation, predictions in regression analysis and, 497
 Interquartile range, 122
 Interval scale, 16
 Influence analysis, 578–580
 Irregular effect, 631

J

Joint probability, 156
 Joint event, 154
 Joint response, 38
 JMP
 Classification and regression trees, 704–705
 Cluster analysis, 706
 Graph builder, 704
 Multidimensional scaling, 706
 Neural networks, 705–706
 treemaps, 702–703
 Judgment sample, 22

K

k-means clustering, 691
 Kruskal-Wallis rank test for differences in c medians, 473–476
 assumptions of, 476
 Kurtosis, 114–115

L

Lagged predictor variable, 647
 Least-squares method in determining simple linear regression, 494–495
 Least-squares trend fitting
 and forecasting, 637–645
 Left-skewed, 114
 Leptokurtic, 115
 Level of confidence, 276
 Level of significance (α), 312
 Levels, 395
 Levene's test
 for homogeneity of variance, 405–406
 Linear regression. *See* Simple linear regression
 Linear relationship, 493
 Linear trend model, 637–638
 Logarithms, rules for, 717
 Logarithmic transformation, 605
 Logistic regression, 573–576
 Logworth statistic, 685

M

Main effects, 423
 Main effects plot, 400
 Managing the Managing Ashland MultiComm Services, 30, 82, 144, 213, 245–246, 269, 301–302, 341, 383–384, 437, 482–483, 536, 587, 668
 Marascuilo procedure, 458–460
 Marginal probability, 157, 167
 Margin of error, 25
 Matched samples, 359
 Mathematical model, 195
 McNemar test, 477
 Mean, 102–104
 of the binomial distribution, 200
 confidence interval estimation for, geometric, 106–107
 of hypergeometric distribution, 207
 population, 127
 sample size determination for, 290
 sampling distribution of, 251–261
 standard error of, 253
 unbiased property of, 251
 Mean absolute deviation, 656
 Mean squares, 397
 Mean Square Among (MSA), 398
 Mean Square A (MSA), 412, 420
 Mean Square B (MSB), 421
 Mean Square Blocks (MSBL), 412
 Mean Square Error (MSE), 412, 421
 Mean Square Interaction (MSAB), 421
 Mean Square Total (MST), 398
 Mean Square Within (MSW), 398
 Measurement
 types of scales, 15–17
 Measurement error, 25
 Median, 104–105
 Microsoft Excel,
 Absolute and relative cell references, 723
 Autocorrelation, 540
 autoregressive modeling, 670–671
 bar charts, 89
 Bayes' theorem, 183
 basic probabilities, 183
 binomial probabilities, 216
 bins, 45
 boxplots, 147
 bullet graphs, 701
 cells, 6
 cell means plot, 444
 cell references, 722
 central tendency, 145
 chart formatting, 727
 checking for and applying Excel updates, 738–739
 checklist for using, 6
 chi-square tests for contingency tables, 486–487
 coefficient of variation, 146
 computing conventions, 7
 confidence interval estimate for the difference between the means of two independent groups, 387
 confidence interval for the mean, 304
 confidence interval for the proportion, 305
 configuring Excel security for add-ins, 739–740
 contingency tables, 86–87
 correlation coefficient, 147
 counting rules, 183
 covariance, 147
 covariance of a probability distribution, 215
 creating histograms for discrete probability distributions, 728
 creating and copying worksheets, 10
 cross-classification table, 86–87
 cumulative percentage distribution, 88–89
 cumulative percentage polygon, 93–94
 descriptive statistics, 145–147
 drilldown, 703
 dummy variables, 591
 entering data, 8–9
 entering array formulas, 724
 entering formulas into worksheets, 723
 establishing the variable type, 33
 expected value, 215
 exponential probabilities, 248
 exponential smoothing, 669
 FAQs, 765–766
 frequency distribution, 87–88
 functions, 723–724
 F test for the ratio of two variances, 390
 Gauges, 700
 Geometric mean, 145
 Getting ready to use, 738
 Guide workbooks, 737
 Histogram, 92
 Hypergeometric probabilities, 216
 Kruskal-Wallis test, 488
 least-squares trend fitting, 670
 Levene test, 441–442
 Logistic regression, 592
 Marascuilo procedure, 487
 moving averages, 669
 multidimensional contingency tables, 94–95
 multiple regression, 589–591
 mean absolute deviation, 671
 model building, 626
 new function names, 760
 normal probabilities, 247
 normal probability plot, 247–248

- one-tail tests, 344
- one-way analysis of variance, 440
- opening workbooks, 9
- ordered array, 87
- quartiles, 146
- Paired *t* test, 388
- Pareto chart, 91
- Pasting with Paste Special, 724
- Percentage distribution, 88–89
- Percentage polygon, 93
- pie chart, 89
- PivotTables, 86
- Poisson probabilities, 216
- Pooled-variance *t* test, 386
- Population parameters, 146
- portfolio expected return, 215
- prediction interval, 540
- preparing and using data,
- printing worksheets, 10
- probability, 183
- probability distribution for a discrete random variable, 215
- quadratic regression, 625
- randomized block, 442
- range, 146
- recalculation, 722–723
- recoding, 33
- relative frequency distribution, 88–89
- residual analysis, 539–540, 590–591
- sample size determination, 305
- sampling distributions, 270
- saving workbooks, 9
- scatter plot, 94
- seasonal data, 671
- selecting cell ranges for charts, 727–728
- separate-variance *t* test, 387
- side-by-side bar chart, 91
- simple linear regression, 538–539
- simple random samples, 33–34
- skill set needed, 8
- slicers, 703
- sparklines, 700
- special note for Office 365 users, 739
- standard deviation, 146
- stem-and-leaf display, 91
- summary tables, 85–86
- t* test for the mean (α unknown), 344
- templates, 8, 723
- time-series plot, 94
- transformations, 625
- treemaps, 701
- two-way analysis of variance, 443
- Tukey-Kramer multiple comparisons, 441
- understanding nonstatistical functions, 762–763
- useful keyboard shortcuts, 759
- variance, 146
- variance inflationary factor (VIF),
- verifying formulas and worksheets, 760
- Wilcoxon rank sum test, 488
- Workbooks, 6
- Worksheet entries and references, 722
- Worksheets, 6
- Worksheet formatting, 724–726
- Z test for the difference between two proportions, 389
- Z test for the mean (α known), 343
- Z scores, 146
- Z test for the proportion, 344
- Midspread, 122
- Minitab
 - autoregressive modeling, 673
 - bar chart, 96
 - best-subsets regression, 628
 - binomial probabilities, 217
 - boxplot, 149
 - chi-square tests for contingency tables, 489
 - collinearity, 627
 - confidence interval for the mean, 306
 - confidence interval for the proportion, 306
 - contingency table, 96
 - correlation coefficient, 150
 - counting rules, 184
 - covariance, 149
 - creating and copying worksheets, 12
 - cross-tabulation table, 96
 - cumulative percentage polygon, 99–100
 - descriptive statistics, 148–149
 - dummy variables, 594
 - entering data, 11
 - establishing the variable type, 34
 - exponential probabilities, 249
 - exponential smoothing, 672
 - F* test for the difference between variances, 392–393
 - FAQs, 766
 - histogram, 98–99
 - geometric mean, 148
 - hypergeometric probabilities, 218
 - Influence analysis, 595
 - Kruskal-Wallis test, 490
 - least-squares trend fitting, 673
 - Levene test, 445
 - logistic regression, 595
 - main effects plot, 444
 - model building, 627
 - moving averages, 672
 - multidimensional contingency tables, 100
 - multiple regression, 592–593
 - normal probabilities, 248
 - normal probability plot, 249
 - one-tail tests, 345
 - one-way analysis of variance, 444–445
 - opening worksheets and projects, 11–12
 - ordered array, 96
 - percentage polygon, 99
 - paired *t* test, 391–392
 - Pareto plot, 97
 - pie chart, 97
 - Poisson probabilities, 217–218
 - probability distribution for a discrete random variable,
 - printing worksheets, 12
 - project, 6
 - quadratic regression, 626
 - randomized block design, 445–446
 - recoding variables, 35
 - residual analysis, 541
 - saving worksheets, 11–12
 - sampling distributions, 271
 - sample size, 307

- saving worksheets and projects, 11–12
 scatter plot, 100
 seasonal data, 673
 side-by-side bar chart, 97
 simple linear regression, 541
 simple random samples, 35
 stacked data, 96
 stem-and-leaf display, 98
 stepwise regression, 627
 summary table, 95–96
 t test for the difference between two means, 391
 t test for the mean (μ unknown), 345
 three-dimensional plot, 592
 time-series plot, 100
 transforming variables, 627
 Tukey-Kramer procedure, 445
 two-way ANOVA, 446
 unstacked data, 96
 variance inflationary factors, 627
 Wilcoxon rank sum test, 489
 Z test for the mean (μ known), 345
 Z test for the difference between two proportions, 392
 Z test for the proportion, 345–346
 Mixed effects models, 431
 Mode, 105–106
 Models. *See* Multiple regression models
 More Descriptive Choices Follow-up, 144, 246, 303, 385, 438, 484–485, 624
 Mountain States Potato Company case, 623, 699
 Moving averages, 632–634
 Multidimensional contingency tables, 68–69
 Multidimensional scaling, 693
 Multiple comparisons, 402
 Multiple regression models, 544
 Adjusted r , 550
 best-subsets approach to, 612–613
 coefficient of multiple determination in, 549–550, 602
 coefficients of partial determination in, 562
 collinearity in, 608–609
 confidence interval estimates for the slope in, 555–556
 dummy-variable models in, 564–566
 ethical considerations in, 618
 interpreting slopes in, 545
 interaction terms, 566–567
 with k independent variables, 545
 model building, 609–611
 model validation, 616
 net regression coefficients, 547
 partial F -test statistic in, 558–561
 pitfalls in, 618
 predicting the dependent variable Y , 547–548
 quadratic, 597–601
 residual analysis for, 553–554
 stepwise regression approach to, 611–612
 testing for significance of, 551
 testing portions of, 558–562
 testing slopes in, 555–556
 transformation in, 605–606
 variance inflationary factors in, 608
 Multilayer perceptrons, 688
 Multiplication rule, 168
 Mutually exclusive events, 157
 Mystatlab course outline,
 Accessing, 729
- N**
- Net regression coefficient, 547
 Neural networks, 688
 Nominal scale, 15–16
 Nonparametric methods, 467
 Nonprobability sample, 21
 Nonresponse bias, 25
 Nonresponse error, 25
 Normal approximation to the binomial distribution, 242
 Normal distribution, 220
 cumulative standardized, 223
 properties of, 220
 Normal probabilities
 calculating, 222–230
 Normal probability density function, 222
 Normal probability plot, 235
 constructing, 235
 Normality assumption, 405
 Null hypothesis, 309
 Numerical descriptive measures
 coefficient of correlation, 133–134
 measures of central tendency, variation, and shape, 102–116
 from a population, 127–131
 Numerical variables, 15
 Organizing, 42–49
 Visualizing, 58–62, 65–67
- O**
- Observed frequency, 449
 Odds ratio, 573
 Ogive, 61–62
 One-tail tests, 328
 null and alternative hypotheses in, 328
 One-way analysis of variance (ANOVA),
 assumptions, 405
 F test for differences in more than two means, 398
 F test statistic, 398
 Levene's test for homogeneity of variance, 405–406
 summary table, 399
 Tukey-Kramer procedure, 402–404
 Online resources, 729
 Operational definitions, 14
 Ordered array, 42
 Ordinal scale, 15–16
 Organize, 2, 14
 Outliers, 113
 Overall F test, 551
- P**
- Paired t test, 360–365
 Parameter, 19
 Pareto chart, 53–55
 Pareto principle, 53
 Parsimony, 610
 principle of, 656
 Partial F -test statistic, 558–561
 PDF files, 737
 Percentage distribution, 45–47
 Percentage polygon, 60–61
 Percentiles, 121
 Permutation, 175

PHStat2

Autocorrelation, 540
 bar chart, 89
 basic probabilities, 183
 best subsets regression, 626
 binomial probabilities, 215–216
 boxplot, 147
 cell means plot, 444
 chi-square (χ^2) test for contingency tables, 486–487
 confidence interval
 for the mean (σ known), 304
 for the mean (σ unknown), 304
 for the difference between two means, 387
 for the mean value, 540
 for the proportion, 305
 contingency tables, 86
 covariance of a probability distribution, 215
 cumulative percentage distributions, 87–88
 cumulative polygons, 93–94
 exponential probabilities, 248
 FAQs, 764–765
 Files, 737
 F test for ratio of two variances, 390
 frequency distributions, 87
 histograms, 91–92
 hypergeometric probabilities, 216
 installing, 737
 Kruskal-Wallis test, 488
 kurtosis, 146
 Levene's test, 441
 Logistic regression, 592
 Marascuilo procedure, 486–487
 Mean, 145
 Median, 145
 Mode, 145
 Model building, 626
 multiple regression, 589–592
 normal probabilities, 247
 normal probability plot, 247
 one-way ANOVA, 440
 one-way tables, 85
 one-tail tests, 344
 opening, 740
 paired t test, 388
 Pareto chart, 90
 Percentage distribution, 87–88
 Percentage polygon, 93–94
 pie chart, 89
 Poisson probabilities, 216
 pooled-variance t test, 386
 portfolio expected return, 215
 portfolio risk, 215
 prediction interval, 540
 quartiles, 146
 randomized block design, 442
 Random sampling, 34
 Residual analysis, 539, 590
 Sample size determination,
 for the mean, 305
 for the proportion, 305
 sampling distributions, 270
 scatter plot, 94
 separate-variance t test, 387

side-by-side bar chart, 90
 simple linear regression, 538
 simple probability, 183
 simple random samples, 34
 skewness, 146
 stacked data, 87
 standard deviation, 146
 stem-and-leaf display, 91
 stepwise regression, 626
 summary tables, 85
 t test for the mean (σ unknown), 343
 two-way ANOVA, 443
 Tukey-Kramer procedure, 441
 Unstacked data, 87
 Wilcoxon rank sum test, 488
 Z test for the mean (σ known), 343
 Z test for the difference in two proportions, 389
 Z test for the proportion, 344
 Pie chart, 52–53
 PivotTables, 68–69
 Platykurtic, 115
 Point estimate, 273
 Poisson distribution, 202
 calculating probabilities, 203–204
 properties of, 202
 Polygons, 93–94
 cumulative percentage, 93–94
 Pooled-variance t test, 348–353
 Population(s), 19
 Population mean, 127, 252
 Population standard deviation, 128, 252
 Population variance, 128
 Portfolio, 191
 Portfolio expected return, 192
 Portfolio risk, 192
 Power of a test, 313
 Practical significance, 336–337
 Prediction interval estimate, 524–525
 Prediction line, 494
 Predictive analytics, 675
 Prescriptive analytics, 675
 Primary data source, 18
 Probability, 152
 a priori, 153
 Bayes' theorem for, 169
 conditional, 161–163
 empirical, 153
 ethical issues and, 177
 joint, 156
 marginal, 157
 simple, 155
 subjective, 153
 Probability density function, 220
 Probability distribution function, 195
 Probability distribution for discrete random variable, 186
 Probability sample, 21
 Processing elements, 688
 Proportions, 45
 chi-square (χ^2) test for differences between two, 448–454
 chi-square (χ^2) test for differences in more than two, 455–458
 confidence interval estimation for, 287–289
 sample size determination for, 292
 sampling distribution of, 262–264

Z test for the difference between two, 371–372
Z test of hypothesis for, 332–335
 p th-order autoregressive model, 648
 p -value, 317
 p -value approach, 317–319

Q

Quadratic regression, 597–601
 Quadratic trend model, 639–640
 Qualitative forecasting methods, 630
 Qualitative variable, 14
 Quantitative forecasting methods, 630
 Quantitative variable, 15
 Quartiles, 120
 Quantile-quantile plot, 235

R

Random effect, 631
 Random effects models, 431
 Randomized block design 410–416
 Randomness and independence, 405
 Random numbers, table of, 22, 742–743
 Range, 107–108
 interquartile, 122
 Ratio scale, 17
 Recoded variable, 20–21
 Rectangular distribution, 237
 Region of nonrejection, 311
 Region of rejection, 311
 Regression analysis. *See* Multiple regression models; Simple linear regression
 Regression coefficients, 495
 Regression trees, 686
 Relative frequency, 45–46
 Relative frequency distribution, 45–47
 Relevant range, 497
 Repeated measurements, 359
 Replicates, 419
 Residual analysis, 507, 655
 Residual plots
 in detecting autocorrelation, 511–512
 in evaluating equal variance, 510
 in evaluating linearity, 508
 in evaluating normality, 509
 in multiple regression, 553–554
 Residuals, 507
 Resistant measures, 123
 Response variable, 493
 Right-skewed, 114
 Robust, 351

S

Sample, 19
 Sample mean, 102
 Sample proportion, 287
 Sample standard deviation, 108–109
 Sample variance, 108
 Sample size determination
 for mean, 290
 for proportion, 292
 Sample space, 154
 Samples,
 cluster, 23
 convenience, 21
 judgment, 22
 nonprobability, 21
 probability, 21
 simple random, 22
 stratified, 23
 systematic, 23
 Sampling
 from nonnormally distributed populations, 257–261
 from normally distributed populations, 254–257
 with replacement, 22
 without replacement, 22
 Sampling distributions, 251
 of the mean, 251–261
 of the proportion, 262–264
 Sampling error, 25, 276
 Scale
 interval, 16
 nominal, 15
 ordinal, 15
 ratio, 17
 Scatter diagram, 492
 Scatter plot, 65–66, 492
 Seasonal effect, 631
 Secondary data source, 18
 Selection bias, 25
 Separate-variance *t* test for differences in two means, 354–356
 Shape, 114
 Side-by-side bar chart, 55
 Simple event, 153
 Simple linear regression,
 assumptions in, 507
 avoiding pitfalls in, 529
 coefficient of determination in, 503–504
 coefficients in, 495
 computations in, 497–499
 Durbin-Watson statistic, 512–514
 equations in, 495
 estimation of mean values and prediction of individual values, 523–525
 inferences about the slope and correlation coefficient, 515–520
 least-squares method in, 494–495
 pitfalls in, 527
 residual analysis, 507–510
 standard error of the estimate in, 505–506
 sum of squares in, 502–503
 Simple probability, 155
 Simple random sample, 22
 Single linkage, 692
 Skewness, 114–115
 Slicers, 679–680
 Slope, 497
 inferences about, 516–519
 interpreting, in multiple regression, 545
 Solver add-in,
 Checking for presence, 741
 Sources of data, 18
 Sparklines, 676
 Spread, 107
 Square-root transformation, 605
 Stacked data, 49
 Standard deviation, 108–109
 of binomial distribution, 200
 of discrete random variable, 188

- of hypergeometric distribution, 207
 of population, 128
 of sum of two random variables, 191
 Standard error of the estimate, 505–506
 Standard error of the mean, 253
 Standard error of the proportion, 263
 Standardized normal random variable, 222
 Statistic, 3
 Statistics, 3, 19
 descriptive, 4
 inferential, 4
 Statistical inference, 4
 Statistical package, 6
 Statistical symbols, 721
 Stem-and-leaf display, 58
 Stepwise regression
 approach to model building, 611–612
 Strata, 23
 Stratified sample, 23
 Stress statistic, 694
 Structured data, 19
 Studentized deleted residuals, t_i , 579
 Studentized range distribution,
 tables, 754–755
 Student's t distribution, 279
 Student tips, 3, 14, 19, 37, 39, 44, 46, 58, 105, 109, 112, 120, 152, 153, 154, 158, 162, 186, 190, 195, 198, 222, 224, 225, 253, 262, 273, 278, 287, 309, 311, 314, 315, 317, 322, 328, 332, 348, 349, 360, 367, 374, 395, 396, 397, 400, 402, 405, 420, 449, 450, 459, 462, 467, 468, 473, 495, 496, 504, 505, 508, 546, 547, 550, 551, 553, 562, 565, 567, 574, 597, 599, 602, 605, 614, 632, 641, 648, 652, 653, 678, 683, 691, 692
 Subjective probability, 153
 Summary table, 38
 Summation notation, 718–720
 Sum of squares, 108
 Sum of squares among blocks (*SSBL*), 411
 Sum of squares among groups (*SSA*), 397
 Sum of squares due to factor A (*SSA*), 419
 Sum of squares due to factor B (*SSB*), 420
 Sum of squares due to regression (*SSR*), 503
 Sum of squares of error (*SSE*), 411, 420, 503
 Sum of squares to interaction (*SSAB*), 420
 Sum of squares total (*SST*), 395, 396, 411, 420, 502
 Sum of squares within groups (*SSW*), 397
 SureValue Convenience Stores, 303, 342, 384, 438, 483, 623
 Survey errors, 24–26
 Symmetrical, 114
 Systematic sample, 23
- T**
- Tableau public
 Bullet graphs, 701–702
 Treemaps, 703
 Tables
 chi-square, 748
 contingency, 39
 Control chart factors, 757
 Durbin-Watson, 756
 F distribution, 749–752
 for categorical data, 37–40
 cumulative standardized normal distribution, 744–745
 of random numbers, 22, 742–743
 standardized normal distribution, 758
 Studentized range, 754–755
 summary, 38
 t distribution, 746–747
 Wilcoxon rank sum, 753
 t distribution, properties of, 280
 Test statistic, 311
 Tests of hypothesis
 Chi-square (χ^2) test for differences
 between c proportions, 455–458
 between two proportions, 448–454
 Chi-square (χ^2) test of independence, 461–466
 F test for the ratio of two variances, 374–377
 F test for the regression model, 560
 F test for the slope, 517–518
 Kruskal-Wallis rank test for differences in c medians, 473–476
 Levene test, 405–406
 Paired t test, 360–365
 pooled-variance t test, 348–353
 separate-variance t test for differences in two means, 354–356
 t test for the correlation coefficient, 519–520
 t test for the mean (σ unknown), 321–324
 t test for the slope, 516–517, 555
 Wilcoxon rank sum test for differences in two medians, 467–471
 Z test for the mean (σ known), 314
 Z test for the difference between two proportions, 367–371
 Z test for the proportion, 332–335
 Think About This, 172, 231, 356, 664
 Third quartile, 120
 Times series, 630
 Time-series forecasting
 autoregressive model, 647–653
 choosing an appropriate forecasting model, 655–657
 component factors of classical multiplicative, 630–631
 exponential smoothing in, 634–636
 least-squares trend fitting and forecasting, 637–645
 moving averages in, 632–634
 seasonal data, 658–663
 Times series plot, 66–67
 Total variation, 396, 411, 419
 Training data, 689
 Transformation formula, 222
 Transformations in regression models
 logarithmic, 605–606
 square-root, 605
 Treemap, 677–678
 Trend, 631
 t test for a correlation coefficient, 519–520
 t test for the mean (σ unknown), 321–324
 t test for the slope, 516–517, 555
 Tukey-Kramer multiple comparison procedure, 402–404
 Tukey multiple comparison procedure, 415–416
 Two-factor factorial design, 418
 Two-sample tests of hypothesis for numerical data,
 F tests for differences in two variances, 374–377
 Paired t test, 360–365
 t tests for the difference in two means, 348–356
 Wilcoxon rank sum test for differences in two medians, 467–471
 Two-tail test, 314
 Two-way analysis of variance
 cell means plot, 426
 factorial design, 418
 interpreting interaction effects, 426–427

multiple comparisons, 424–425
 testing for factor and interaction effects, 421–422
 Type I error, 312
 Type II error, 312

U

Unbiased, 251
 Unexplained variation or error sum of squares (SSE), 502–503
 Uniform probability distribution, 237
 mean, 237
 standard deviation, 238
 Unstacked data, 49
 Unstructured data, 19

V

Variables, 3
 categorical, 14
 continuous, 15
 discrete, 15
 dummy, 564–566
 numerical, 15
 Variance inflationary factor (VIF), 608
 Variance,
 of discrete random variable, 187
F-test for the ratio of two, 374–377
 Levene's test for homogeneity of, 405–406
 of the sum of two random variables, 191
 population, 128
 sample, 108–109
 Variation, 102
 Venn diagrams, 155
 Visual Explorations, 737

descriptive statistics, 117
 normal distribution, 231
 sampling distributions, 261
 simple linear regression, 499–500
 using, 741

Visualize, 2, 14
 Visualizations,
 Guidelines for constructing, 74

W

Wald statistic, 576
 Ward's minimum variance method, 692
 Width of class interval, 43
 Wilcoxon rank sum test
 for differences in two medians, 467–471
 Tables, 746–747
 Wilcoxon signed ranks test, 478
 Within-group variation, 397

X

Y
 Y intercept b_0 , 494
Z
 Z scores, 113
 Z test,
 for the difference between two proportions, 371–372
 for the mean (σ known), 314
 for the proportion, 332–335

BREAKTHROUGH

To improving results

MyStatLabTM

for Business Statistics

MyStatLab is a course management system that provides engaging learning experiences and delivers proven results while helping students succeed. Tools are embedded which make it easy to integrate statistical software into the course. And, MyStatLab comes from an experienced partner with educational expertise and an eye on the future.

Tutorial Exercises

MyStatLab homework and practice exercises correlated to the exercises in the textbook are generated algorithmically, giving students unlimited opportunity for practice and mastery. MyStatLab grades homework and provides feedback and guidance.

The screenshot shows a homework problem titled "Homework Ch. 13" with a score of 0% (0 of 4 pts) and 0 of 4 complete. The problem asks to construct a scatter plot of Franchise Value vs. Annual Revenue for 15 major sport teams. It provides four options labeled A, B, C, and D, each showing a different scatter plot. Below the plots are three questions:

- Construct a scatter plot. Choose the correct graph below.
- Use the least-squares method to determine the regression coefficients b_0 and b_1 .
 $b_0 = -486.8837$
 $b_1 = 4.7062$
(Round to four decimal places as needed.)
- Interpret the meaning of b_0 and b_1 . Choose the correct answer below.
 A. A practical interpretation of the Y-intercept, b_0 , is not meaningful because no sports franchise is going to have a revenue of zero. The slope, b_1 , implies that for each increase of 1 million dollars in annual revenue, the franchise value is expected to increase by the value of b_1 , in millions of dollars.

At the bottom right of the interface, there are buttons for "Similar Exercise" and "Save".

Help Me Solve This breaks the problem into manageable steps. Students enter answers along the way.

View an Example walks students through a problem similar to the one assigned.

Textbook links to the appropriate section in the eText.

Tech Help is a suite of Technology Tutorial videos that show how to perform statistical calculations using popular software.

Powerful Homework and Test Manager

Create, import, and manage online homework assignments, quizzes, and tests that are automatically graded, allowing you to spend less time grading and more time teaching. Thousands of high-quality and algorithmic exercises of all types and difficulty levels are available to meet the needs of students with diverse mathematical backgrounds.

Ready-to-Go Courses

Ready-to-Go Courses make it even easier for first-time users to start using MyStatLab. With the help of experienced instructors, these courses include pre-made assignments that you can alter at any time.

Adaptive Learning

An Adaptive Study Plan serves as a personalized tutor for your students. When enabled, Knewton in MyStatLab monitors student performance and provides personalized recommendations. It gathers information about learning preferences and is continuously adaptive, guiding students through the Study Plan one objective at a time.

Integrated Statistical Software

Copy our data sets, from the eText and the MyStatLab questions, into software such as StatCrunch, Minitab, Excel, and more. Students have access to support tools—videos, Study Cards, and manuals for select titles—to learn how to use statistical software.

The screenshot shows the 'Tools for Success' section of the MyStatLab interface. On the left is a vertical navigation bar with links like Course Home, Homework, Quizzes & Tests, Study Plan, Gradebook, StatCrunch, Chapter Contents, Tools for Success (which is highlighted in blue), Multimedia Library, Pearson Tutor Services, Discussions, Course Tools, and Instructor Resources. The main content area has three sections: 'Business Statistics Technology Tutorial Videos' (with links to Excel 2010, Excel 2010 with XLSTAT, Minitab, and JMP videos); 'Business Statistics Technology Study Cards' (with links to various software study cards); and 'Graphing Calculator Help' (with a link to a TI calculator tutorial).

MyStatLab includes web-based statistical software, StatCrunch, within the online assessment platform so that students can analyze data sets from exercises and the text. In addition, MyStatLab includes access to www.StatCrunch.com, the full web-based program where users can access thousands of shared data sets, create and conduct online surveys, perform complex analyses using the powerful statistical software, and generate compelling reports.

Engaging Video Resources

- [Business Insight Videos](#) are 10 engaging videos showing managers at top companies using statistics in their everyday work. Assignable questions encourage discussion.
- [StatTalk Videos](#), hosted by fun-loving statistician Andrew Vickers, demonstrate important statistical concepts through interesting stories and real-life events. This series of 24 videos includes available assessment questions and an instructor's guide.

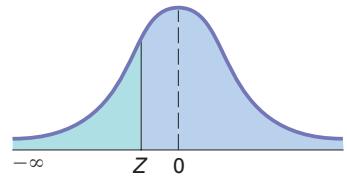
PHStat™ (access code required)

PHStat is a statistics add-in for Microsoft Excel that simplifies the task of operating Excel, creating real Excel worksheets that use in-worksheet calculations. Download PHStat by visiting www.pearsonhighered.com/phstat or through a link in MyStatLab's Tools for Success, access code required.

This book features PHStat version 4 which is compatible with all current Microsoft Windows and (Mac) OS X Excel versions.

The Cumulative Standardized Normal Distribution

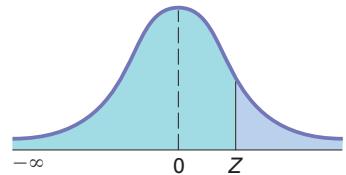
Entry represents area under the cumulative standardized normal distribution from $-\infty$ to Z



Z	Cumulative Probabilities									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-6.0	0.000000001									
-5.5	0.000000019									
-5.0	0.000000287									
-4.5	0.000003398									
-4.0	0.000031671									
-3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003
-3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
-3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
-3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
-3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
-3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
-3.3	0.00048	0.00047	0.00045	0.00043	0.00042	0.00040	0.00039	0.00038	0.00036	0.00035
-3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
-3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
-3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00103	0.00100
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2388	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2482	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

The Cumulative Standardized Normal Distribution (continued)

Entry represents area under the cumulative standardized normal distribution from $-\infty$ to Z



Cumulative Probabilities