FOR THE AP® EXAM

The PRACTICE of STATISTICS

FIFTH EDITION

STARNES • TABOR • YATES • MOORE

AP® is a trademark registered by the College Board®, which was not involved in the production of, and does not endorse, this product.

This page intentionally left blank

Using *The Practice of Statistics*, Fifth Edition, for Advanced Placement (AP®) Statistics (The percents in parentheses reflect coverage on the AP® exam.)	
Topic Outline for AP® Statistics from the College Board's *AP® Statistics Course Description*	***The Practice of Statistics,* 5th ed.** Chapter and Section references
I. Exploring data: describing patterns and departures from patterns (20%–30%)	
A. Constructing and interpreting graphical displays of distributions of univariate data (dotplot, stemplot, histogram, cumulative frequency plot)	**Dotplot, stemplot, histogram 1.2; Cumulative frequency plot 2.1**
1. Center and spread	1.2
2. Clusters and gaps	1.2
3. Outliers and unusual features	1.2
4. Shape	1.2
B. Summarizing distributions of univariate data	**1.3 and 2.1**
1. Measuring center: median, mean	1.3
2. Measuring spread: range, interquartile range, standard deviation	1.3
3. Measuring position: quartiles, percentiles, standardized scores (z-scores)	Quartiles 1.3; percentiles and z-scores 2.1
4. Using boxplots	1.3
5. The effect of changing units on summary measures	2.1
C. Comparing distributions of univariate data (dotplots, back-to-back stemplots, parallel boxplots)	**Dotplots and stemplots 1.2; boxplots 1.3**
1. Comparing center and spread	1.2 and 1.3
2. Comparing clusters and gaps	1.2 and 1.3
3. Comparing outliers and unusual features	1.2 and 1.3
4. Comparing shape	1.2 and 1.3
D. Exploring bivariate data	**Chapter 3 and Section 12.2**
1. Analyzing patterns in scatterplots	3.1
2. Correlation and linearity	3.1
3. Least-squares regression line	3.2
4. Residual plots, outliers, and influential points	3.2
5. Transformations to achieve linearity: logarithmic and power transformations	12.2
E. Exploring categorical data	**Sections 1.1, 5.2, 5.3**
1. Frequency tables and bar charts	1.1 (we call them bar graphs)
2. Marginal and joint frequencies for two-way tables	Marginal 1.1; joint 5.2
3. Conditional relative frequencies and association	1.1 and 5.3
4. Comparing distributions using bar charts	1.1
II. Sampling and experimentation: planning and conducting a study (10%–15%)	
A. Overview of methods of data collection	**Sections 4.1 and 4.2**
1. Census	4.1
2. Sample survey	4.1
3. Experiment	4.2
4. Observational study	4.2
B. Planning and conducting surveys	**Section 4.1**
1. Characteristics of a well-designed and well-conducted survey	4.1
2. Populations, samples, and random selection	4.1
3. Sources of bias in sampling and surveys	4.1
4. Sampling methods, including simple random sampling, stratified random sampling, and cluster sampling	4.1
C. Planning and conducting experiments	**Section 4.2**
1. Characteristics of a well-designed and well-conducted experiment	4.2
2. Treatments, control groups, experimental units, random assignments, and replication	4.2
3. Sources of bias and confounding, including placebo effect and blinding	4.2
4. Completely randomized design	4.2
5. Randomized block design, including matched pairs design	4.2
D. Generalizability of results and types of conclusions that can be drawn from observational studies, experiments, and surveys	**Section 4.3**

Using *The Practice of Statistics*, Fifth Edition, for Advanced Placement (AP®) Statistics (The percents in parentheses reflect coverage on the AP® exam.)	
Topic Outline for AP® Statistics from the College Board's *AP® Statistics Course Description*	***The Practice of Statistics,* 5th ed.** Chapter and Section references
III. Anticipating patterns: exploring random phenomena using probability and simulation (20%–30%)	
A. Probability	**Chapters 5 and 6**
1. Interpreting probability, including long-run relative frequency interpretation	5.1
2. "Law of large numbers" concept	5.1
3. Addition rule, multiplication rule, conditional probability, and independence	Addition rule 5.2; other three topics 5.3
4. Discrete random variables and their probability distributions, including binomial and geometric	Discrete 6.1; Binomial and geometric 6.3
5. Simulation of random behavior and probability distributions	5.1
6. Mean (expected value) and standard deviation of a random variable, and linear transformation of a random variable	Mean and standard deviation 6.1; Linear transformation 6.2
B. Combining independent random variables	**Section 6.2**
1. Notion of independence versus dependence	6.2
2. Mean and standard deviation for sums and differences of independent random variables	6.2
C. The Normal distribution	**Section 2.2**
1. Properties of the Normal distribution	2.2
2. Using tables of the Normal distribution	2.2
3. The Normal distribution as a model for measurements	2.2
D. Sampling distributions	**Chapter 7; Sections 8.3, 10.1, 10.2, 11.1**
1. Sampling distribution of a sample proportion	7.2
2. Sampling distribution of a sample mean	7.3
3. Central limit theorem	7.3
4. Sampling distribution of a difference between two independent sample proportions	10.1
5. Sampling distribution of a difference between two independent sample means	10.2
6. Simulation of sampling distributions	7.1
7. t distribution	8.3
8. Chi-square distribution	11.1
IV. Statistical inference: estimating population parameters and testing hypotheses (30%–40%)	
A. Estimation (point estimators and confidence intervals)	**Chapter 8 plus parts of Sections 9.3, 10.1, 10.2, 12.1**
1. Estimating population parameters and margins of error	8.1
2. Properties of point estimators, including unbiasedness and variability	8.1
3. Logic of confidence intervals, meaning of confidence level and confidence intervals, and properties of confidence intervals	8.1
4. Large-sample confidence interval for a proportion	8.2
5. Large-sample confidence interval for a difference between two proportions	10.1
6. Confidence interval for a mean	8.3
7. Confidence interval for a difference between two means (unpaired and paired)	Paired 9.3; unpaired 10.2
8. Confidence interval for the slope of a least-squares regression line	12.1
B. Tests of significance	**Chapters 9 and 11 plus parts of Sections 10.1, 10.2, 12.1**
1. Logic of significance testing, null and alternative hypotheses; P-values; one-and two-sided tests; concepts of Type I and Type II errors; concept of power	9.1; power in 9.2
2. Large-sample test for a proportion	9.2
3. Large-sample test for a difference between two proportions	10.1
4. Test for a mean	9.3
5. Test for a difference between two means (unpaired and paired)	Paired 9.3; unpaired 10.2
6. Chi-square test for goodness of fit, homogeneity of proportions, and independence (one- and two-way tables)	Chapter 11
7. Test for the slope of a least-squares regression line	12.1

For the **AP**® Exam

The Practice of Statistics

FIFTH EDITION

AP® is a trademark registered by the College Board, which was not involved in the production of, and does not endorse, this product.

Publisher: Ann Heath
Assistant Editor: Enrico Bruno
Editorial Assistant: Matt Belford
Development Editor: Donald Gecewicz
Executive Marketing Manager: Cindi Weiss
Photo Editor: Cecilia Varas
Photo Researcher: Julie Tesser
Art Director: Diana Blume
Cover Designers: Diana Blume, Rae Grant
Text Designer: Patrice Sheridan
Cover Image: Joseph Devenney/Getty Images
Senior Project Editor: Vivien Weiss
Illustrations: Network Graphics
Production Manager: Susan Wein
Composition: Preparé
Printing and Binding: RR Donnelley

TI-83™, TI-84™, TI-89™, and TI-Nspire screen shots are used with permission of the publisher:
© 1996, Texas Instruments Incorporated.
TI-83™, TI-84™, TI-89™, and TI-Nspire Graphic Calculators are registered trademarks of Texas Instruments Incorporated.
Minitab is a registered trademark of Minitab, Inc.
Microsoft© and Windows© are registered trademarks of the Microsoft Corporation in the United States and other countries.
Fathom Dynamic Statistics is a trademark of Key Curriculum, a McGraw-Hill Education Company.

M&M'S is a registered trademark of Mars, Incorporated and its affiliates. This trademark is used with permission. Mars, Incorporated is not associated with Macmillan Higher Education. Images printed with permission of Mars, Incorporated.

Library of Congress Control Number: 2013949503
ISBN-13: 978-1-4641-0873-0
ISBN-10: 1-4641-0873-0
© 2015, 2012, 2008, 2003, 1999 by W. H. Freeman and Company

First printing 2014

All rights reserved

Printed in the United States of America

W. H. Freeman and Company
41 Madison Avenue
New York, NY 10010
Houndmills, Basingstoke RG21 6XS, England
www.whfreeman.com

For the **AP**® Exam

FIFTH EDITION

The Practice of Statistics

Daren S. Starnes
The Lawrenceville School

Josh Tabor
Canyon del Oro High School

Daniel S. Yates
Statistics Consultant

David S. Moore
Purdue University

bfw Freeman

W. H. Freeman and Company/BFW
New York

Contents

About the Authors

DAREN S. STARNES is Mathematics Department Chair and holds the Robert S. and Christina Seix Dow Distinguished Master Teacher Chair in Mathematics at The Lawrenceville School near Princeton, New Jersey. He earned his MA in Mathematics from the University of Michigan and his BS in Mathematics from the University of North Carolina at Charlotte. Daren is also an alumnus of the North Carolina School of Science and Mathematics. Daren has led numerous one-day and weeklong AP® Statistics institutes for new and experienced AP® teachers, and he has been a Reader, Table Leader, and Question Leader for the AP® Statistics exam since 1998. Daren is a frequent speaker at local, state, regional, national, and international conferences. For two years, he served as coeditor of the Technology Tips column in the NCTM journal *The Mathematics Teacher*. From 2004 to 2009, Daren served on the ASA/NCTM Joint Committee on the Curriculum in Statistics and Probability (which he chaired in 2009). While on the committee, he edited the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) pre-K–12 report and coauthored (with Roxy Peck) *Making Sense of Statistical Studies*, a capstone module in statistical thinking for high school students. Daren is also coauthor of the popular text *Statistics Through Applications*, First and Second Editions.

JOSH TABOR has enjoyed teaching general and AP® Statistics to high school students for more than 18 years, most recently at his alma mater, Canyon del Oro High School in Oro Valley, Arizona. He received a BS in Mathematics from Biola University, in La Mirada, California. In recognition of his outstanding work as an educator, Josh was named one of the five finalists for Arizona Teacher of the Year in 2011. He is a past member of the AP® Statistics Development Committee (2005–2009), as well as an experienced Table Leader and Question Leader at the AP® Statistics Reading. Each year, Josh leads one-week AP® Summer Institutes and one-day College Board workshops around the country and frequently speaks at local, national, and international conferences. In addition to teaching and speaking, Josh has authored articles in *The Mathematics Teacher*, *STATS Magazine*, and *The Journal of Statistics Education*. He is the author of the *Annotated Teacher's Edition* and *Teacher's Resource Materials* for *The Practice of Statistics* 4e and 5e, along with the *Solutions Manual* for *The Practice of Statistics* 5e. Combining his love of statistics and love of sports, Josh teamed with Christine Franklin to write *Statistical Reasoning in Sports*, an innovative textbook for on-level statistics courses.

DANIEL S. YATES taught AP® Statistics in the Electronic Classroom (a distance-learning facility) affiliated with Henrico County Public Schools in Richmond, Virginia. Prior to high school teaching, he was on the mathematics faculty at Virginia Tech and Randolph-Macon College. He has a PhD in Mathematics Education from Florida State University. Dan received a College Board/Siemens Foundation Advanced Placement Teaching Award in 2000.

DAVID S. MOORE is Shanti S. Gupta Distinguished Professor of Statistics (Emeritus) at Purdue University and was 1998 President of the American Statistical Association. David is an elected fellow of the American Statistical Association and of the Institute of Mathematical Statistics and an elected member of the International Statistical Institute. He has served as program director for statistics and probability at the National Science Foundation. He is the author of influential articles on statistics education and of several leading textbooks.

Content Advisory Board and Supplements Team

Jason Molesky, Lakeville Area Public School, Lakeville, MN
Media Coordinator—Worked Examples, PPT lectures, Strive for a 5 Guide

Jason has served as an AP® Statistics Reader and Table Leader since 2006. After teaching AP® Statistics for eight years and developing the FRAPPY system for AP® exam preparation, Jason moved into administration. He now serves as the Director of Program Evaluation and Accountability, overseeing the district's research and evaluation, continuous improvement efforts, and assessment programs. Jason also provides professional development to statistics teachers across the United States and maintains the "Stats Monkey" Web site, a clearinghouse for AP® Statistics resources.

Tim Brown, The Lawrenceville School, Lawrenceville, NJ
Content Advisor, Test Bank, TRM Tests and Quizzes

Tim first piloted an AP® Statistics course the year before the first exam was administered. He has been an AP® Reader since 1997 and a Table Leader since 2004. He has taught math and statistics at The Lawrenceville School since 1982 and currently holds the Bruce McClellan Distinguished Teaching Chair.

Doug Tyson, Central York High School, York, PA
Exercise Videos, Learning Curve

Doug has taught mathematics and statistics to high school and undergraduate students for 22 years. He has taught AP® Statistics for 7 years and served as an AP® Reader for 4 years. Doug is the co-author of a curriculum module for the College Board, conducts student review sessions around the country, and gives workshops on teaching statistics.

Paul Buckley, Gonzaga College High School, Washington, DC
Exercise Videos

Paul has taught high school math for 20 years and AP® Statistics for 12 years. He has been an AP® Statistics Reader for six years and helps to coordinate the integration of new Readers (Acorns) into the Reading process. Paul has presented at Conferences for AP®, NCTM, NCEA (National Catholic Education Association) and JSEA (Jesuit Secondary Education Association).

Leigh Nataro, Moravian Academy, Bethlehem, PA
Technology Corner Videos

Leigh has taught AP® Statistics for nine years and has served as an AP® Statistics Reader for the past four years. She enjoys the challenge of writing multiple-choice questions for the College Board for use on the AP® Statistics exam. Leigh is a National Board Certified Teacher in Adolescence and Young Adulthood Mathematics and was previously named a finalist for the Presidential Award for Excellence in Mathematics and Science Teaching in New Jersey.

Ann Cannon, Cornell College, Mount Vernon, IA
Content Advisor, Accuracy Checker

Ann has served as Reader, Table Leader, and Question Leader for the AP® Statistics exam for the past 13 years. She has taught introductory statistics at the college level for 20 years and is very active in the Statistics Education Section of the American Statistical Association, serving on the Executive Committee for two 3-year terms. She is co-author of *STAT2: Building Models for a World of Data* (W. H. Freeman and Company).

Michel Legacy, Greenhill School, Dallas, TX
Content Advisor, Strive for a 5 Guide

Michael is a past member of the AP® Statistics Development Committee (2001–2005) and a former Table Leader at the Reading. He currently reads the Alternate Exam and is a lead teacher at many AP® Summer Institutes. Michael is the author of the 2007 College Board AP® Statistics Teacher's Guide and was named the Texas 2009–2010 AP® Math/Science Teacher of the Year by the Siemens Corporation.

James Bush, Waynesburg University, Waynesburg, PA
Learning Curve, Media Reviewer

James has taught introductory and advanced courses in Statistics for over 25 years. He is currently a Professor of Mathematics at Waynesburg University and is the recipient of the Lucas Hathaway Teaching Excellence Award. James has served as an AP® Statistics Reader for the past seven years and conducts many AP® Statistics preparation workshops.

Beth Benzing, Strath Haven High School, Wallingford/Swarthmore School District, Wallingford, PA
Activities Videos

Beth has taught AP® Statistics for 14 years and has served as a Reader for the AP® Statistics exam for the past four years. She serves as Vice President on the board for the regional affiliate for NCTM in the Philadelphia area and is a moderator for an on-line course, *Teaching Statistics with Fathom*.

Heather Overstreet, Franklin County High School, Rocky Mount, VA
TI-Nspire Technology Corners

Heather has taught AP® Statistics for nine years and has served as an AP® Statistics Reader for the past six years. While working with Virginia Advanced Study Strategies, a program for promoting AP® math, science, and English courses in Virginia High Schools, she led many AP® Statistics Review Sessions and served as a Laying the Foundation trainer of teachers of pre-AP® math classes.

Acknowledgments

Teamwork: It has been the secret to the successful evolution of *The Practice of Statistics* (*TPS*) through its fourth and fifth editions. We are indebted to each and every member of the team for the time, energy, and passion that they have invested in making our collective vision for *TPS* a reality.

To our team captain, Ann Heath, we offer our utmost gratitude. Managing a revision project of this scope is a Herculean task! Ann has a knack for troubleshooting thorny issues with an uncanny blend of forthrightness and finesse. She assembled an all-star cast to collaborate on the fifth edition of *TPS* and trusted each of us to deliver an excellent finished product. We hope you'll agree that the results speak for themselves. Thank you, Ann, for your unwavering support, patience, and friendship throughout the production of these past two editions.

Truth be told, we had some initial reservations about adding a Development Editor to our team starting with the fourth edition of *TPS*. Don Gecewicz quickly erased our doubts. His keen mind and sharp eye were evident in the many probing questions and insightful suggestions he offered at all stages of the project. Thanks, Don, for your willingness to push the boundaries of our thinking.

Working behind the scenes, Enrico Bruno busily prepared the manuscript chapters for production. He did yeoman's work in ensuring that the changes we intended were made as planned. For the countless hours that Enrico invested sweating the small stuff, we offer him our sincere appreciation.

We are deeply grateful to Patrice Sheridan and to Diana Blume for their aesthetic contributions to the eye-catching design of *TPS* 5e. Our heartfelt thanks also go to Vivien Weiss and Susan Wein at W. H. Freeman for their skillful oversight of the production process. Patti Brecht did a superb job copyediting a very complex manuscript.

A special thank you goes to our good friends on the high school sales and marketing staff at Bedford, Freeman, and Worth (BFW) Publishers. We feel blessed to have such enthusiastic professionals on our extended team. In particular, we want to thank our chief cheerleader, Cindi Weiss, for her willingness to promote *The Practice of Statistics* at every opportunity.

We cannot say enough about the members of our Content Advisory Board and Supplements Team. This remarkable group is a veritable who's who of the AP® Statistics community. We'll start with Ann Cannon, who once again reviewed the statistical content of every chapter. Ann also checked the solutions to every exercise in the book. What a task! More than that, Ann offered us sage advice about virtually everything between the covers of *TPS* 5e. We are so grateful to Ann for all that she has done to enhance the quality of *The Practice of Statistics* over these past two editions.

Jason Molesky, aka "Stats Monkey," greatly expanded his involvement in the fifth edition by becoming our Media Coordinator in addition to his ongoing roles as author of the *Strive for a 5 Guide* and creator of the PowerPoint presentations for the book. Jason also graciously loaned us his Free Response AP® Problem, Yay! (FRAPPY!) concept for use in *TPS* 5e. We feel incredibly fortunate to have such a creative, energetic, and deeply thoughtful person at the helm as the media side of our project explodes in many new directions at once.

Jason is surrounded by a talented media team. Doug Tyson and Paul Buckley have expertly recorded the worked exercise videos. Leigh Nataro has produced screencasts for the Technology Corners on the TI-83/84, TI-89, and TI-Nspire. Beth Benzing followed through on her creative suggestion to produce "how to" screencasts for teachers to accompany the Activities in the book. James Bush capably served as our expert reviewer for all of the media elements and is now partnering with Doug Tyson on a new Learning Curve media component. Heather Overstreet once

again compiled the TI-Nspire Technology Corners in Appendix B. We wish to thank the entire media team for their many contributions.

Tim Brown, my Lawrenceville colleague, has been busy creating many new, high-quality assessment items for the fifth edition. The fruits of his labors are contained in the revised *Test Bank* and in the quizzes and tests that are part of the *Teacher's Resource Materials*. We are especially thankful that Tim was willing to compose an additional cumulative test for Chapters 2 through 12.

We offer our continued thanks to Michael Legacy for composing the superb questions in the Cumulative AP® Practice Tests and the *Strive for a 5 Guide*. Michael's expertise as a two-time former member of the AP® Statistics Test Development Committee is invaluable to us.

Although Dan Yates and David Moore both retired several years ago, their influence lives on in *TPS* 5e. They both had a dramatic impact on our thinking about how statistics is best taught and learned through their pioneering work as textbook authors. Without their early efforts, *The Practice of Statistics* would not exist!

Thanks to all of you who reviewed chapters of the fourth and fifth editions of *TPS* and offered your constructive suggestions. The book is better as a result of your input.

—Daren Starnes and Josh Tabor

A final note from Daren: It has been a privilege for me to work so closely with Josh Tabor on *TPS* 5e. He is a gifted statistics educator; a successful author in his own right; and a caring parent, colleague, and friend. Josh's influence is present on virtually every page of the book. He also took on the thankless task of revising all of the solutions for the fifth edition in addition to updating his exceptional *Annotated Teacher's Edition*. I don't know how he finds the time and energy to do all that he does!

The most vital member of the *TPS* 5e team for me is my wonderful wife, Judy. She has read page proofs, typed in data sets, and endured countless statistical conversations in the car, in airports, on planes, in restaurants, and on our frequent strolls. If writing is a labor of love, then I am truly blessed to share my labor with the person I love more than anyone else in the world. Judy, thank you so much for making the seemingly impossible become reality time and time again. And to our three sons, Simon, Nick, and Ben—thanks for the inspiration, love, and support that you provide even in the toughest of times.

A final note from Josh: When Daren asked me to join the *TPS* team for the fourth edition, I didn't know what I was getting into. Having now completed another full cycle with *TPS* 5e, I couldn't have imagined the challenge of producing a textbook—from the initial brainstorming sessions to the final edits on a wide array of supplementary materials. In all honesty, I still don't know the full story. For taking on all sorts of additional tasks—managing the word-by-word revisions to the text, reviewing copyedits and page proofs, encouraging his co-author, and overseeing just about everything—I owe my sincere thanks to Daren. He has been a great colleague and friend throughout this process.

I especially want to thank the two most important people in my life. To my wife Anne, your patience while I spent countless hours working on this project is greatly appreciated. I couldn't have survived without your consistent support and encouragement. To my daughter Jordan, I look forward to being home more often and spending less time on my computer when I am there. I also look forward to when you get to use *TPS* 7e in about 10 years. For now, we have a lot of fun and games to catch up on. I love you both very much.

Acknowledgments

Fifth Edition Survey Participants and Reviewers

Blake Abbott, Bishop Kelley High School, Tulsa, OK
Maureen Bailey, Millcreek Township School District, Erie, PA
Kevin Bandura, Lincoln County High School, Stanford, KY
Elissa Belli, Highland High School, Highland, IN
Jeffrey Betlan, Yough School District, Herminie, PA
Nancy Cantrell, Macon County Schools, Franklin, NC
Julie Coyne, Center Grove HS, Greenwood, IN
Mary Cuba, Linden Hall, Lititz, PA
Tina Fox, Porter-Gaud School, Charleston, SC
Ann Hankinson, Pine View, Osprey, FL
Bill Harrington, State College Area School District, State College, PA
Ronald Hinton, Pendleton Heights High School, Pendleton, IN
Kara Immonen, Norton High School, Norton, MA
Linda Jayne, Kent Island High School, Stevensville, MD
Earl Johnson, Chicago Public Schools, Chicago, IL
Christine Kashiwabara, Mid-Pacific Institute, Honolulu, HI
Melissa Kennedy, Holy Names Academy, Seattle, WA
Casey Koopmans, Bridgman Public Schools, Bridgman, MI
David Lee, SPHS, Sun Prairie, WI
Carolyn Leggert, Hanford High School, Richland, WA
Jeri Madrigal, Ontario High School, Ontario, CA
Tom Marshall, Kents Hill School, Kents Hill, ME
Allen Martin, Loyola High School, Los Angeles, CA
Andre Mathurin, Bellarmine College Preparatory, San Jose, CA
Brett Mertens, Crean Lutheran High School, Irvine, CA
Sara Moneypenny, East High School, Denver, CO
Mary Mortlock, The Harker School, San Jose, CA
Mary Ann Moyer, Hollidaysburg Area School District, Hollidaysburg, PA
Howie Nelson, Vista Murrieta HS, Murrieta, CA
Shawnee Patry, Goddard High School, Wichita, KS
Sue Pedrick, University High School, Hartford, CT
Shannon Pridgeon, The Overlake School, Redmond, WA
Sean Rivera, Folsom High, Folsom, CA
Alyssa Rodriguez, Munster High School, Munster, IN
Sheryl Rodwin, West Broward High School, Pembroke Pines, FL
Sandra Rojas, Americas HS, El Paso, TX
Christine Schneider, Columbia Independent School, Boonville, MO
Amanda Schneider, Battle Creek Public Schools, Charlotte, MI
Steve Schramm, West Linn High School, West Linn, OR
Katie Sinnott, Revere High School, Revere, MA
Amanda Spina, Valor Christian High School, Highlands Ranch, CO
Julie Venne, Pine Crest School, Fort Lauderdale, FL
Dana Wells, Sarasota High School, Sarasota, FL
Luke Wilcox, East Kentwood High School, Grand Rapids, MI
Thomas Young, Woodstock Academy, Putnam, CT

Fourth Edition Focus Group Participants and Reviewers

Gloria Barrett, Virginia Advanced Study Strategies, Richmond, VA
David Bernklau, Long Island University, Brookville, NY
Patricia Busso, Shrewsbury High School, Shrewsbury, MA
Lynn Church, Caldwell Academy, Greensboro, NC
Steven Dafilou, Springside High School, Philadelphia, PA
Sandra Daire, Felix Varela High School, Miami, FL
Roger Day, Pontiac High School, Pontiac, IL
Jared Derksen, Rancho Cucamonga High School, Rancho Cucamonga, CA
Michael Drozin, Munroe Falls High School, Stow, OH
Therese Ferrell, I. H. Kempner High School, Sugar Land, TX
Sharon Friedman, Newport High School, Bellevue, WA
Jennifer Gregor, Central High School, Omaha, NE
Julia Guggenheimer, Greenwich Academy, Greenwich, CT
Dorinda Hewitt, Diamond Bar High School, Diamond Bar, CA
Dorothy Klausner, Bronx High School of Science, Bronx, NY
Robert Lochel, Hatboro-Horsham High School, Horsham, PA
Lynn Luton, Duchesne Academy of the Sacred Heart, Houston, TX
Jim Mariani, Woodland Hills High School, Greensburgh, PA
Stephen Miller, Winchester Thurston High School, Pittsburgh, PA
Jason Molesky, Lakeville Area Public Schools, Lakeville, MN
Mary Mortlock, Harker School, San Jose, CA
Heather Nichols, Oak Creek High School, Oak Creek, WI
Jamis Perrett, Texas A&M University, College Station, TX
Heather Pessy, Mount Lebanon High School, Pittsburgh, PA
Kathleen Petko, Palatine High School, Palatine, IL
Todd Phillips, Mills Godwin High School, Richmond, VA
Paula Schute, Mount Notre Dame High School, Cincinnati, OH
Susan Stauffer, Boise High School, Boise, ID
Doug Tyson, Central York High School, York, PA
Bill Van Leer, Flint High School, Oakton, VA
Julie Verne, Pine Crest High School, Fort Lauderdale, FL

Steve Willot, Francis Howell North High School, St. Charles, MO
Jay C. Windley, A. B. Miller High School, Fontana, CA

Reviewers of previous editions:

Christopher E. Barat, Villa Julie College, Stevenson, MD
Jason Bell, Canal Winchester High School, Canal Winchester, OH
Zack Bigner, Elkins High School, Missouri City, TX
Naomi Bjork, University High School, Irvine, CA
Robert Blaschke, Lynbrook High School, San Jose, CA
Alla Bogomolnaya, Orange High School, Pepper Pike, OH
Andrew Bowen, Grand Island Central School District, Grand Island, NY
Jacqueline Briant, Bishop Feehan High School, Attleboro, MA
Marlys Jean Brimmer, Ridgeview High School, Bakersfield, CA
Floyd E. Brown, Science Hill High School, Johnson City, TN
James Cannestra, Germantown High School, Germantown, WI
Joseph T. Champine, King High School, Corpus Christi, TX
Jared Derksen, Rancho Cucamonga High School, Rancho Cucamonga, CA
George J. DiMundo, Coast Union High School, Cambria, CA
Jeffrey S. Dinkelmann, Novi High School, Novi, MI
Ronald S. Dirkse, American School in Japan, Tokyo, Japan
Cynthia L. Dishburger, Whitewater High School, Fayetteville, GA
Michael Drake, Springfield High School, Erdenheim, PA
Mark A. Fahling, Gaffney High School, Gaffney, SC
David Ferris, Noblesville High School, Noblesville, IN
David Fong, University High School, Irvine, CA
Terry C. French, Lake Braddock Secondary School, Burke, VA
Glenn Gabanski, Oak Park and River Forest High School, Oak Park, IL
Jason Gould, Eaglecrest High School, Centennial, CO
Dr. Gene Vernon Hair, West Orange High School, Winter Garden, FL
Stephen Hansen, Napa High School, Napa, CA
Katherine Hawks, Meadowcreek High School, Norcross, GA
Gregory D. Henry, Hanford West High School, Hanford, CA
Duane C. Hinders, Foothill College, Los Altos Hills, CA
Beth Howard, Saint Edwards, Sebastian, FL
Michael Irvin, Legacy High School, Broomfield, CO
Beverly A. Johnson, Fort Worth Country Day School, Fort Worth, TX
Matthew L. Knupp, Danville High School, Danville, KY
Kenneth Kravetz, Westford Academy, Westford, MA
Lee E. Kucera, Capistrano Valley High School, Mission Viejo, CA
Christina Lepi, Farmington High School, Farmington, CT
Jean E. Lorenson, Stone Ridge School of the Sacred Heart, Bethesda, MD
Thedora R. Lund, Millard North High School, Omaha, NE
Philip Mallinson, Phillips Exeter Academy, Exeter, NH
Dru Martin, Ramstein American High School, Ramstein, Germany
Richard L. McClintock, Ticonderoga High School, Ticonderoga, NY
Louise McGuire, Pattonville High School, Maryland Heights, MO
Jennifer Michaelis, Collins Hill High School, Suwanee, GA
Dr. Jackie Miller, Ohio State University
Jason M. Molesky, Lakeville South High School, Lakeville, MN
Wayne Nirode, Troy High School, Troy, OH
Heather Pessy, Mount Lebanon High School, Pittsburgh, PA
Sarah Peterson, University Preparatory Academy, Seattle, WA
Kathleen Petko, Palatine High School, Palatine, IL
German J. Pliego, University of St. Thomas
Stoney Pryor, A&M Consolidated High School, College Station, TX
Judy Quan, Alameda High School, Alameda, CA
Stephanie Ragucci, Andover High School, Andover, MA
James M. Reeder, University School, Hunting Valley, OH
Joseph Reiken, Bishop Garcia Diego High School, Santa Barbara, CA
Roger V. Rioux, Cape Elizabeth High School, Cape Elizabeth, ME
Tom Robinson, Kentridge Senior High School, Kent, WA
Albert Roos, Lexington High School, Lexington, MA
Linda C. Schrader, Cuyahoga Heights High School, Cuyahoga Heights, OH
Daniel R. Shuster, Royal High School, Simi Valley, CA
David Stein, Paint Branch High School, Burtonsville, MD
Vivian Annette Stephens, Dalton High School, Dalton, GA
Charles E. Taylor, Flowing Wells High School, Tucson, AZ
Reba Taylor, Blacksburg High School, Blacksburg, VA
Shelli Temple, Jenks High School, Jenks, OK
David Thiel, Math/Science Institute, Las Vegas, NV
William Thill, Harvard-Westlake School, North Hollywood, CA
Richard Van Gilst, Westminster Christian Academy, St. Louis, MO
Joseph Robert Vignolini, Glen Cove High School, Glen Cove, NY
Ira Wallin, Elmwood Park Memorial High School, Elmwood Park, NJ
Linda L. Wohlever, Hathaway Brown School, Shaker Heights, OH

To the Student

Statistical Thinking and You

The purpose of this book is to give you a working knowledge of the big ideas of statistics and of the methods used in solving statistical problems. Because data always come from a real-world context, doing statistics means more than just manipulating data. *The Practice of Statistics* (*TPS*), Fifth Edition, is full of data. Each set of data has some brief background to help you understand what the data say. We deliberately chose contexts and data sets in the examples and exercises to pique your interest.

TPS 5e is designed to be easy to read and easy to use. This book is written by current high school AP® Statistics teachers, for high school students. We aimed for clear, concise explanations and a conversational approach that would encourage you to read the book. We also tried to enhance both the visual appeal and the book's clear organization in the layout of the pages.

Be sure to take advantage of all that *TPS* 5e has to offer. You can learn a lot by reading the text, but you will develop deeper understanding by doing Activities and Data Explorations and answering the Check Your Understanding questions along the way. The walkthrough guide on pages xiv–xx gives you an inside look at the important features of the text.

You learn statistics best by doing statistical problems. This book offers many different types of problems for you to tackle.

- **Section Exercises** include paired odd- and even-numbered problems that test the same skill or concept from that section. There are also some multiple-choice questions to help prepare you for the AP® exam. Recycle and Review exercises at the end of each exercise set involve material you studied in previous sections.
- **Chapter Review Exercises** consist of free-response questions aligned to specific learning objectives from the chapter. Go through the list of learning objectives summarized in the Chapter Review and be sure you can say "I can do that" to each item. Then prove it by solving some problems.
- The **AP® Statistics Practice Test** at the end of each chapter will help you prepare for in-class exams. Each test has 10 to 12 multiple-choice questions and three free-response problems, very much in the style of the AP® exam.
- Finally, the **Cumulative AP® Practice Tests** after Chapters 4, 7, 10, and 12 provide challenging, cumulative multiple-choice and free-response questions like ones you might find on a midterm, final, or the AP® Statistics exam.

The main ideas of statistics, like the main ideas of any important subject, took a long time to discover and take some time to master. The basic principle of learning them is to be persistent. Once you put it all together, statistics will help you make informed decisions based on data in your daily life.

TPS and AP® Statistics

The Practice of Statistics (*TPS*) was the first book written specifically for the Advanced Placement (AP®) Statistics course. Like the previous four editions, *TPS* 5e is organized to closely follow the AP® Statistics Course Description. Every item on the College Board's "Topic Outline" is covered thoroughly in the text. Look inside the front cover for a detailed alignment guide. The few topics in the book that go beyond the AP® syllabus are marked with an asterisk (*).

Most importantly, *TPS* 5e is designed to prepare you for the AP® Statistics exam. The entire author team has been involved in the AP® Statistics program since its early days. We have more than 80 years' combined experience teaching introductory statistics and more than 30 years' combined experience grading the AP® exam! Two of us (Starnes and Tabor) have served as Question Leaders for several years, helping to write scoring rubrics for free-response questions. Including our Content Advisory Board and Supplements Team (page vii), we have two former Test Development Committee members and 11 AP® exam Readers.

TPS 5e will help you get ready for the AP® Statistics exam throughout the course by:

- **Using terms, notation, formulas, and tables consistent with those found on the AP® exam.** Key terms are shown in bold in the text, and they are defined in the Glossary. Key terms also are cross-referenced in the Index. See page F-1 to find "Formulas for the AP® Statistics Exam" as well as Tables A, B, and C in the back of the book for reference.
- **Following accepted conventions from AP® exam rubrics when presenting model solutions.** Over the years, the scoring guidelines for free-response questions have become fairly consistent. We kept these guidelines in mind when writing the solutions that appear throughout *TPS* 5e. For example, the four-step State-Plan-Do-Conclude process that we use to complete inference problems in Chapters 8 through 12 closely matches the four-point AP® scoring rubrics.
- **Including AP® Exam Tips in the margin where appropriate.** We place exam tips in the margins and in some Technology Corners as "on-the-spot" reminders of common mistakes and how to avoid them. These tips are collected and summarized in Appendix A.
- **Providing hundreds of AP®-style exercises throughout the book.** We even added a new kind of problem just prior to each Chapter Review, called a FRAPPY (Free Response AP® Problem, Yay!). Each FRAPPY gives you the chance to solve an AP®-style free-response problem based on the material in the chapter. After you finish, you can view and critique two example solutions from the book's Web site (www.whfreeman.com/tps5e). Then you can score your own response using a rubric provided by your teacher.

Turn the page for a tour of the text. See how to use the book to realize success in the course and on the AP® exam.

READ THE TEXT and use the book's features to help you grasp the big ideas.

Read the LEARNING OBJECTIVES at the beginning of each section. Focus on mastering these skills and concepts as you work through the chapter.

3.1 Scatterplots and Correlation

WHAT YOU WILL LEARN By the end of the section, you should be able to:

- Identify explanatory and response variables in situations where one variable helps to explain or influences the other.
- Make a scatterplot to display the relationship between two quantitative variables.
- Describe the direction, form, and strength of a relationship displayed in a scatterplot and identify outliers in a scatterplot.
- Interpret the correlation.
- Understand the basic properties of correlation, including how the correlation is influenced by outliers.
- Use technology to calculate correlation.
- Explain why association does not imply causation.

Scan the margins for the purple notes, which represent the "voice of the teacher" giving helpful hints for being successful in the course.

Often, using the regression line to make a prediction for $x = 0$ is an extrapolation. That's why the y intercept isn't always statistically meaningful.

DEFINITION: Extrapolation

Extrapolation is the use of a regression line for prediction far outside the interval of values of the explanatory variable x used to obtain the line. Such predictions are often not accurate.

Few relationships are linear for all values of the explanatory variable. *Don't make predictions using values of x that are much larger or much smaller than those that actually appear in your data.*

Take note of the green DEFINITION boxes that explain important vocabulary. Flip back to them to review key terms and their definitions.

Watch for CAUTION ICONS. They alert you to common mistakes that students make.

Look for the boxes with the blue bands. Some explain how to make graphs or set up calculations while others recap important concepts.

HOW TO MAKE A SCATTERPLOT

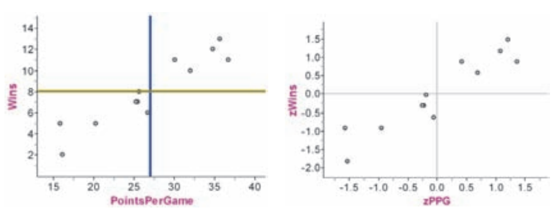
1. Decide which variable should go on each axis.
2. Label and scale your axes.
3. Plot individual data values.

Make connections and deepen your understanding by reflecting on the questions asked in THINK ABOUT IT passages.

THINK ABOUT IT

What does correlation measure? The Fathom screen shots below provide more detail. At the left is a scatterplot of the SEC football data with two lines added—a vertical line at the group's mean points per game and a horizontal line at the mean number of wins of the group. Most of the points fall in the upper-right or lower-left "quadrants" of the graph. That is, teams with above-average points per game tend to have above-average numbers of wins, and teams with below-average points per game tend to have numbers of wins that are below average. This confirms the positive association between the variables.

Below on the right is a scatterplot of the standardized scores. To get this graph, we transformed both the x- and the y-values by subtracting their mean and dividing by their standard deviation. As we saw in Chapter 2, standardizing a data set converts the mean to 0 and the standard deviation to 1. That's why the vertical and horizontal lines in the right-hand graph are both at 0.



Notice that all the products of the standardized values will be positive—not surprising, considering the strong positive association between the variables. What if there was a negative association between two variables? Most of the points would be in the upper-left and lower-right "quadrants" and their z-score products would be negative, resulting in a negative correlation.

Read the AP® EXAM TIPS. They give advice on how to be successful on the AP® exam.

AP® EXAM TIP The formula sheet for the AP® exam uses different notation for these equations: $b_1 = r\frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1\bar{x}$. That's because the least-squares line is written as $\hat{y} = b_0 + b_1x$. We prefer our simpler versions without the subscripts!

LEARN STATISTICS BY *DOING* STATISTICS

ACTIVITY | Reaching for Chips

MATERIALS:
200 colored chips, including 100 of the same color; large bag or other container

Before class, your teacher will prepare a population o
having the same color (say, red). The parameter is th
chips in the population: $p = 0.50$. In this Activity, y
variability by taking repeated random samples of size

1. After your teacher has mixed the chips thoroughly
should take a sample of 20 chips and note the sampl
When finished, the student should return all the chi
and pass the bag to the next student.

Note: If your class has fewer than 25 students, have s
samples.

2. Each student should record the $\hat{p}$-value in a chart
value on a class dotplot. Label the graph scale from (
spaced 0.05 units apart.

3. Describe what you see: shape, center, spread, and
usual features.

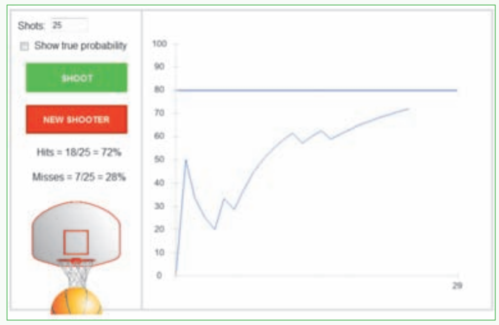
ACTIVITY | I'm a Great Free-Throw Shooter!

MATERIALS:
Computer with Internet access and projection capability

A basketball player claims to make 80% of the free throws that he attempts. We think he might be exaggerating. To test this claim, we'll ask him to shoot some free throws—virtually—using *The Reasoning of a Statistical Test* applet at the book's Web site.

APPLET

1. Go to www.whfreeman.com/tps5e and launch the applet.



2. Set the applet to take 25 shots. Click "Shoot." How many of the 25 shots did the player make? Do you have enough data to decide whether the player's claim is valid?
3. Click "Shoot" again for 25 more shots. Keep doing this until you are convinced *either* that the player makes less than 80% of his shots *or* that the player's claim is true. How large a sample of shots did you need to make your decision?
4. Click "Show true probability" to reveal the truth. Was your conclusion correct?
5. If time permits, choose a new shooter and repeat Steps 2 through 4. Is it easier to tell that the player is exaggerating when his actual proportion of free throws made is closer to 0.8 or farther from 0.8?

Every chapter begins with a hands-on ACTIVITY that introduces the content of the chapter. Many of these activities involve collecting data and drawing conclusions from the data. In other activities, you'll use dynamic applets to explore statistical concepts.

DATA EXPLORATION The SAT essay: Is longer better?

Following the debut of the new SAT Writing test in March 2005, Dr. Les Perelman from the Massachusetts Institute of Technology stirred controversy by reporting, "It appeared to me that regardless of what a student wrote, the longer the essay, the higher the score." He went on to say, "I have never found a quantifiable predictor in 25 years of grading that was anywhere as strong as this one. If you just graded them based on length without ever reading them, you'd be right over 90 percent of the time."[3] The table below shows the data that Dr. Perelman used to draw his conclusions.[4]

Length of essay and score for a sample of SAT essays

Words:	460	422	402	365	357	278	236	201	168	156	133
Score:	6	6	5	5	6	5	4	4	4	3	2
Words:	114	108	100	403	401	388	320	258	236	189	128
Score:	2	1	1	5	6	6	5	4	4	3	2
Words:	67	697	387	355	337	325	272	150	135		
Score:	1	6	6	5	5	4	4	2	3		

Does this mean that if students write a lot, they are guaranteed high scores? Carry out your own analysis of the data. How would you respond to each of Dr. Perelman's claims?

DATA EXPLORATIONS ask you to play the role of data detective. Your goal is to answer a puzzling, real-world question by examining data graphically and numerically.

CHECK YOUR UNDERSTANDING

Identify the explanatory and response variables in each setting.

1. How does drinking beer affect the level of alcohol in people's blood? The legal limit for driving in all states is 0.08%. In a study, adult volunteers drank different numbers of cans of beer. Thirty minutes later, a police officer measured their blood alcohol levels.

2. The National Student Loan Survey provides data on the amount of debt for recent college graduates, their current income, and how stressed they feel about college debt. A sociologist looks at the data with the goal of using amount of debt and income to explain the stress caused by college debt.

CHECK YOUR UNDERSTANDING questions appear throughout the section. They help you to clarify definitions, concepts, and procedures. Be sure to check your answers in the back of the book.

EXAMPLES: Model statistical problems and how to solve them

You will often see explanatory variables called *independent variables* and response variables called *dependent variables*. Because the words "independent" and "dependent" have other meanings in statistics, we won't use them here.

It is easiest to identify explanatory and response variables when we actually specify values of one variable to see how it affects another variable. For instance, to study the effect of alcohol on body temperature, researchers gave several different amounts of alcohol to mice. Then they measured the change in each mouse's body temperature 15 minutes later. In this case, amount of alcohol is the explanatory variable, and change in body temperature is the response variable. When we don't specify the values of either variable but just observe both variables, there may or may not be explanatory and response variables. Whether there are depends on how you plan to use the data.

Read through each EXAMPLE, and then try out the concept yourself by working the FOR PRACTICE exercise in the Section Exercises.

EXAMPLE Linking SAT Math and Critical Reading Scores

Explanatory or response?

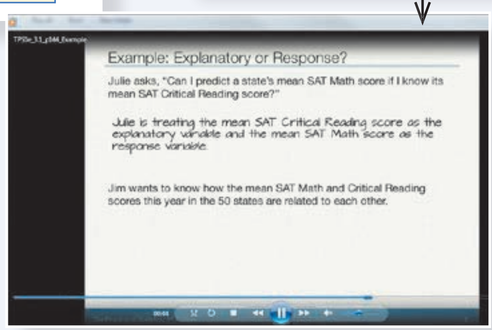
Julie asks, "Can I predict a state's mean SAT Math score if I know its mean SAT Critical Reading score?" Jim wants to know how the mean SAT Math and Critical Reading scores this year in the 50 states are related to each other.

PROBLEM: For each student, identify the explanatory variable and the response variable if possible.

SOLUTION: Julie is treating the mean SAT Critical Reading score as the explanatory variable and the mean SAT Math score as the response variable. Jim is simply interested in exploring the relationship between the two variables. For him, there is no clear explanatory or response variable.

For Practice *Try Exercise* 1

Need extra help? Examples and exercises marked with the PLAY ICON are supported by short video clips prepared by experienced AP® teachers. The video guides you through each step in the example and solution and gives you extra help when you need it.



The red number box next to the exercise directs you back to the page in the section where the model example appears.

1\. **Coral reefs** How sensitive to changes in water temperature are coral reefs? To find out, measure the growth of corals in aquariums where the water temperature is controlled at different levels. Growth is measured by weighing the coral before and after the experiment. What are the explanatory and response variables? Are they categorical or quantitative?

pg 144

EXAMPLE Gesell Scores

STEP 4

Putting it all together

Does the age at which a child begins to talk predict a later score on a test of mental ability? A study of the development of young children recorded the age in months at which each of 21 children spoke their first word and their Gesell Adaptive Score, the result of an aptitude test taken much later.[16] The data appear in the table below, along with a scatterplot, residual plot, and computer output. Should we use a linear model to predict a child's Gesell score from his or her age at first word? If so, how accurate will our predictions be?

Age (months) at first word and Gesell score

CHILD	AGE	SCORE	CHILD	AGE	SCORE	CHILD	AGE	SCORE
1	15	95	8	11	100	15	11	102
2	26	71	9	8	104	16	10	100
3	10	83	10	20	94	17	12	105
4	9	91	11	7	113	18	42	57
5	15	102	12	9	96	19	17	121
6	20	87	13	10	83	20	11	86
7	18	93	14	11	84	21	10	100

4-STEP EXAMPLES: By reading the 4-Step Examples and mastering the special "State-Plan-Do-Conclude" framework, you can develop good problem-solving skills and your ability to tackle more complex problems like those on the AP® exam.

EXERCISES: Practice makes perfect!

Start by reading the SECTION SUMMARY to be sure that you understand the key concepts.

Section 3.2 Summary

- A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. You can use a regression line to **predict** the value of y for any value of x by substituting this x into the equation of the line.
- The **slope** b of a regression line $\hat{y} = a + bx$ is the rate at which the predicted response $\hat{y}$ changes along the line as the explanatory variable x changes. Specifically, b is the *predicted* change in y when x increases by 1 unit.
- The **y intercept** a of a regression line $\hat{y} = a + bx$ is the predicted response $\hat{y}$ when the explanatory variable x equals 0. This prediction is of no statistical use unless x can actually take values near 0.

Practice! Work the EXERCISES assigned by your teacher. Compare your answers to those in the Solutions Appendix at the back of the book. Short solutions to the exercises numbered in red are found in the appendix.

Most of the exercises are paired, meaning that odd- and even-numbered problems test the same skill or concept. If you answer an assigned problem incorrectly, try to figure out your mistake. Then see if you can solve the paired exercise.

Section 3.2 Exercises

35. **What's my line?** You use the same bar of soap to shower each morning. The bar weighs 80 grams when it is new. Its weight goes down by 6 grams per day on average. What is the equation of the regression line for predicting weight from days of use?

36. **What's my line?** An eccentric professor believes that a child with IQ 100 should have a reading test score of 50 and predicts that reading score should increase by 1 point for every additional point of IQ. What is the equation of the professor's regression line for predicting reading score from IQ?

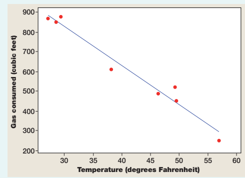
37. **Gas mileage** We expect a car's highway gas mileage to be related to its city gas mileage. Data for all 1198 vehicles in the government's recent *Fuel Economy Guide* give the regression line: predicted highway mpg = 4.62 + 1.109 (city mpg).
(a) What's the slope of this line? Interpret this value in context.
(b) What's the y intercept? Explain why the value of the intercept is not statistically meaningful.
(c) Find the predicted highway mileage for a car that gets 16 miles per gallon in the city.

38. **IQ and reading scores** Data on the IQ test scores and reading test scores for a group of fifth-grade children give the following regression line: predicted reading score = −33.4 + 0.882(IQ sco
(a) What's the slope of this line? Interpret thi context.
(b) What's the y intercept? Explain why the intercept is not statistically meaningful.
(c) Find the predicted reading score for a cl IQ score of 90.

39. **Acid rain** Researchers studying acid rai the acidity of precipitation in a Colorad area for 150 consecutive weeks. Acidity i by pH. Lower pH values show higher ac researchers observed a linear pattern ove They reported that the regression line pl 0.0053(weeks) fit the data well.[19]

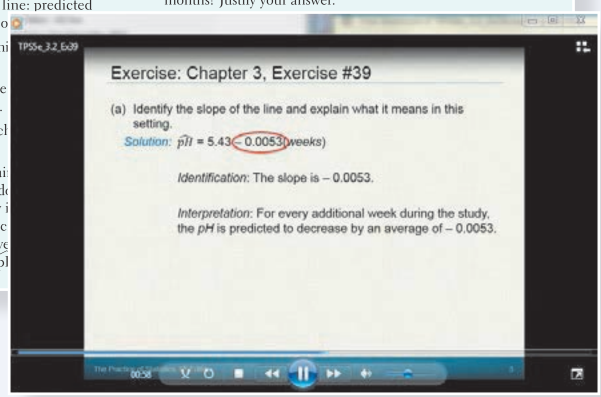
pg 166

in Joan's midwestern home. The figure below shows the original scatterplot with the least-squares line added. The equation of the least-squares line is $\hat{y} = 1425 - 19.87x$.



(a) Identify the slope of the line and explain what it means in this setting.
(b) Identify the y intercept of the line. Explain why it's risky to use this value as a prediction.
(c) Use the regression line to predict the amount of natural gas Joan will use in a month with an average temperature of 30°F.

41. **Acid rain** Refer to Exercise 39. Would it be appropriate to use the regression line to predict pH after 1000 months? Justify your answer.



Look for icons that appear next to selected problems. They will guide you to

- an Example that models the problem.
- videos that provide step-by-step instructions for solving the problem.
- earlier sections on which the problem draws (here, Section 2.2).
- examples with the 4-Step State-Plan-Do-Conclude way of solving problems.

79. **In my Chevrolet** (2.2) The Chevrolet Malibu with a four-cylinder engine has a combined gas mileage of 25 mpg. What percent of all vehicles have worse gas mileage than the Malibu?

STEP 4

67. **Beavers and beetles** Do beavers benefit beetles? Researchers laid out 23 circular plots, each 4 meters in diameter, in an area where beavers were cutting down cottonwood trees. In each plot, they counted the number of stumps from trees cut by beavers and the number of clusters of beetle larvae. Ecologists think that the new sprouts from stumps are more tender than other cottonwood growth, so that beetles prefer them.

pg 185

REVIEW and PRACTICE for quizzes and tests

Chapter Review

Section 3.1: Scatterplots and Correlation

In this section, you learned how to explore the relationship between two quantitative variables. As with distributions of a single variable, the first step is always to make a graph. A scatterplot is the appropriate type of graph to investigate associations between two quantitative variables. To describe a scatterplot, be sure to discuss four characteristics: direction, form, strength, and outliers. The direction of an association might be positive, negative, or neither. The form of an association can be linear or nonlinear. An association is strong if it closely follows a specific form. Finally, outliers are any points that clearly fall outside the pattern of the rest of the data.

The correlation *r* is a numerical summary that describes the direction and strength of a linear association. When $r > 0$, the association is positive, and when $r < 0$, the association is negative. The correlation will always take values between -1 and 1, with $r = -1$ and $r = 1$ indic ing a perfectly linear relationship. Strong linear assoc tions have correlations near 1 or -1, while weak lin relationships have correlations near 0. However, it is possible to determine the form of an association from only the correlation. Strong nonlinear relationships can have a correlation close to 1 or a correlation close to 0, depending on the association. You also learned that outliers can greatly affect the value of the correlation and that correlation does not imply causation. That is, we can't assume that changes in one variable cause changes in the other variable, just because they have a correlation close to 1 or –1.

Section 3.2: Least-Squares Regression

In this section, you learned how to use least-squares regression lines as models for relationships between variables that have a linear association. It is important to understand the difference between the actual data and the model used to describe the data. For example, when you are interpreting the slope of a least-squares regression

Review the CHAPTER SUMMARY to be sure that you understand the key concepts in each section.

What Did You Learn?

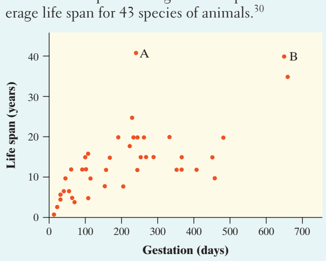
Learning Objective	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
Identify explanatory and response variables in situations where one variable helps to explain or influences the other.	3.1	144	R3.4
Make a scatterplot to display the relationship between two quantitative variables.	3.1	145, 148	R3.4
Describe the direction, form, and strength of a relationship displayed in a scatterplot and recognize outliers in a scatterplot.	3.1	147, 148	R3.1
Interpret the correlation.	3.1	152	R3.3, R3.4
Understand the basic properties of correlation, including how the correlation is influenced by outliers.	3.1	152, 156, 157	R3.1, R3.2
Use technology to calculate correlation.	3.1	Activity on 152, 171	R3.4
Explain why association does not imply causation.	3.1	Discussion on 156, 190	R3.6
Interpret the slope and *y* intercept of a least-squares regression line.	3.2	166	R3.2, R3.4
Use the least-squares regression line to predict *y* for a given *x*. Explain the dangers of extrapolation.	3.2	167, Discussion on 168 (for extrapolation)	R3.2, R3.4, R3.5
Calculate and interpret residuals.	3.2	169	R3.3, R3.4
Explain the concept of least squares.	3.2	Discussion on 169	R3.5
Determine the equation of a least-squares regression line using technology or computer output.	3.2	Technology Corner on 171, 181	R3.3, R3.4
Construct and interpret residual plots to assess whether a linear model is appropriate.	3.2	Discussion on 175, 180	R3.3, R3.4
Interpret the standard deviation of the residuals and r^2 and use these values to assess how well the least-squares regression line models the relationship between two variables.	3.2	180	R3.3, R3.5
		Discussion on 188	R3.1
		183	R3.5

Use the WHAT DID YOU LEARN? table to guide you to model examples and exercises to verify your mastery of each LEARNING OBJECTIVE.

Chapter 3 Chapter Review Exercises

These exercises are designed to help you review the important ideas and methods of the chapter.

R3.1 **Born to be old?** Is there a relationship between the gestational period (time from conception to birth) of an animal and its average life span? The figure shows a scatterplot of the gestational period and average life span for 43 species of animals.[30]

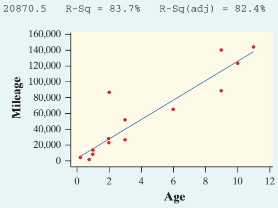


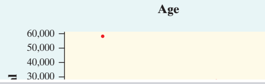
(a) Describe the association shown in the scatterplot.

R3.3 **Stats teachers' cars** A random sample of AP® Statistics teachers was asked to report the age (in years) and mileage of their primary vehicles. A scatterplot of the data, a least-squares regression printout, and a residual plot are provided below.

```
Predictor    Coef     SE Coef     T       P
Constant     3704     8268        0.45    0.662
Age          12188    1492        8.17    0.000

S = 20870.5   R-Sq = 83.7%   R-Sq(adj) = 82.4%
```





Tackle the CHAPTER REVIEW EXERCISES for practice in solving problems that test concepts from throughout the chapter.

and the AP® Exam

Each chapter concludes with an AP® STATISTICS PRACTICE TEST. This test includes about 10 AP®-style multiple-choice questions and 3 free-response questions.

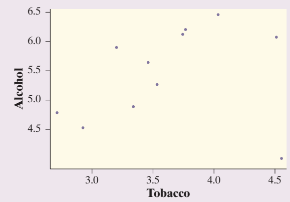
Chapter 3 AP® Statistics Practice Test

Section I: Multiple Choice *Select the best answer for each question.*

T3.1 A school guidance counselor examines the number of extracurricular activities that students do and their grade point average. The guidance counselor says, "The evidence indicates that the correlation between the number of extracurricular activities a student participates in and his or her grade point average is close to zero." A correct interpretation of this statement would be that

(a) active students tend to be students with poor grades, and vice versa.
(b) students with good grades tend to be students who are not involved in many extracurricular activities, and vice versa.
(c) students involved in many extracurricular activities are just as likely to get good grades as bad grades; the same is true for students involved in few extracurricular activities.
(d) there is no linear relationship between number of activ-

alcoholic beverages for each of 11 regions in Great Britain was recorded. A scatterplot of spending on alcohol versus spending on tobacco is shown below. Which of the following statements is true?



(a) The observation (4.5, 6.0) is an outlier.
(b) There is clear evidence of a negative association be-
...ending on alcohol and tobacco.
...ation of the least-squares line for this plot
...e approximately $\hat{y} = 10 - 2x$.
...relation for these data is $r = 0.99$.
...ervation in the lower-right corner of the plot is
...al for the least-squares line.

Cumulative AP® Practice Test 1

Section I: Multiple Choice *Choose the best answer for Questions AP1.1 to AP1.14.*

AP1.1 You look at real estate ads for houses in Sarasota, Florida. Many houses range from $200,000 to $400,000 in price. The few houses on the water, however, have prices up to $15 million. Which of the following statements best describes the distribution of home prices in Sarasota?

(a) The distribution is most likely skewed to the left, and the mean is greater than the median.

AP1.4 For a certain experiment, the available experimental units are eight rats, of which four are female (F1, F2, F3, F4) and four are male (M1, M2, M3, M4). There are to be four treatment groups, A, B, C, and D. If a randomized block design is used, with the experimental units blocked by gender, which of the following assignments of treatments is impossible?

Section II: Free Response Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

AP1.15 The manufacturer of exercise machines for fitness centers has designed two new elliptical machines that are meant to increase cardiovascular fitness. The two machines are being tested on 30 volunteers at a fitness center near the company's headquarters. The volunteers are randomly assigned to one of the machines and use it daily for two months. A measure of cardiovascular fitness is administered at the start of the experiment and the two machines. Note that higher scores indicate larger gains in fitness.

Machine A		Machine B
	0	2
5 4	1	0
8 7 6 3 2 0	2	1 5 9
9 7 4 1 1	3	2 4 8 9

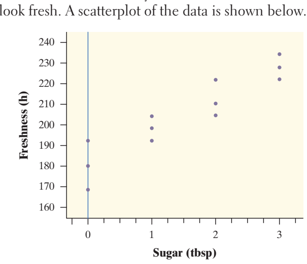
Four CUMULATIVE AP® TESTS simulate the real exam. They are placed after Chapters 4, 7, 10, and 12. The tests expand in length and content coverage from the first through the fourth.

FRAPPY! Free Response AP® Problem, Yay!

The following problem is modeled after actual AP® Statistics exam free response questions. Your task is to generate a complete, concise response in 15 minutes.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

Two statistics students went to a flower shop and randomly selected 12 carnations. When they got home, the students prepared 12 identical vases with exactly the same amount of water in each vase. They put one tablespoon of sugar in 3 vases, two tablespoons of sugar in 3 vases, and three tablespoons of sugar in 3 vases. In the remaining 3 vases, they put no sugar. After the vases were prepared, the students randomly assigned 1 carnation to each vase and observed how many hours each flower continued to look fresh. A scatterplot of the data is shown below.



(a) Briefly describe the association shown in the scatterplot.
(b) The equation of the least-squares regression line for these data is $\hat{y} = 180.8 + 15.8x$. Interpret the slope of the line in the context of the study.
(c) Calculate and interpret the residual for the flower that had 2 tablespoons of sugar and looked fresh for 204 hours.
(d) Suppose that another group of students conducted a similar experiment using 12 flowers, but included different varieties in addition to carnations. Would you expect the value of r^2 for the second group's data to be greater than, less than, or about the same as the value of r^2 for the first group's data? Explain.

After you finish, you can view two example solutions on the book's Web site (www.whfreeman.com/tps5e). Determine whether you think each solution is "complete," "substantial," "developing," or "minimal." If the solution is not complete, what improvements would you suggest to the student who wrote it? Finally, your teacher will provide you with a scoring rubric. Score your response and note what, if anything, you would do differently to improve your own score.

Learn how to answer free-response questions successfully by working the FRAPPY! THE FREE RESPONSE AP® PROBLEM, YAY! that comes just before the Chapter Review in every chapter.

Use TECHNOLOGY to discover and analyze

Use technology as a tool for discovery and analysis. TECHNOLOGY CORNERS give step-by-step instructions for using the TI-83/84 and TI-89 calculator. Instructions for the TI-Nspire are in an end-of-book appendix. HP Prime instructions are on the book's Web site and in the e-Book.

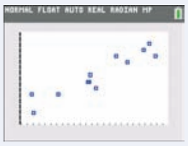
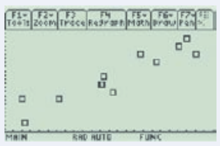
7. TECHNOLOGY CORNER **SCATTERPLOTS ON THE CALCULATOR**

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site

Making scatterplots with technology is much easier than constructing them by hand. We'll use the SEC football data from page 146 to show how to construct a scatterplot on a TI-83/84 or TI-89.

- Enter the data values into your lists. Put the points per game in L1/list1 and the number of wins in L2/list2.
- Define a scatterplot in the statistics plot menu (press F2 on the TI-89). Specify the settings shown below.


- Use ZoomStat (ZoomData on the TI-89) to obtain a graph. The calculator will set the window dimensions automatically by looking at the values in L1/list1 and L2/list2.

Notice that there are no scales on the axes and that the axes are not labeled. If you copy a scatterplot from your calculator onto your paper, make sure that you scale and label the axes.

AP® EXAM TIP If you are asked to make a scatterplot on a free-response question, be sure to label and scale both axes. *Don't* just copy an unlabeled calculator graph directly onto your paper.

You can access video instructions for the Technology Corners through the e-Book or on the book's Web site.

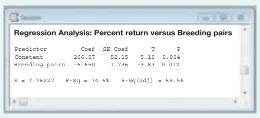
Find the Technology Corners easily by consulting the summary table at the end of each section or the complete table inside the back cover of the book.

3.2 TECHNOLOGY CORNERS

TI-Nspire Instructions in Appendix B; HP Prime instructions on the book's Web site

59. **Merlins breeding** Exercise 13 (page 160) gives data on the number of breeding pairs of merlins in an isolated area in each of seven years and the percent of males who returned the next year. The data show that the percent returning is lower after successful breeding seasons and that the relationship is roughly linear. The figure below shows Minitab regression output for these data.

pg 181



(a) What is the equation of the least-squares regression line for predicting the percent of males that return from the number of breeding pairs? Use the equation to predict the percent of returning males after a season with 30 breeding pairs.

(b) What percent of the year-to-year variation in percent of returning males is accounted for by the straight-line relationship with number of breeding pairs the previous year?

Other types of software displays, including Minitab, Fathom, and applet screen captures, appear throughout the book to help you learn to read and interpret many different kinds of output.

Overview What Is Statistics?

Does listening to music while studying help or hinder learning? If an athlete fails a drug test, how sure can we be that she took a banned substance? Does having a pet help people live longer? How well do SAT scores predict college success? Do most people recycle? Which of two diets will help obese children lose more weight and keep it off? Should a poker player go "all in" with pocket aces? Can a new drug help people quit smoking? How strong is the evidence for global warming?

These are just a few of the questions that statistics can help answer. But what is statistics? And why should you study it?

Statistics Is the Science of Learning from Data

Data are usually numbers, but they are not "just numbers." *Data are numbers with a context.* The number 10.5, for example, carries no information by itself. But if we hear that a family friend's new baby weighed 10.5 pounds at birth, we congratulate her on the healthy size of the child. The context engages our knowledge about the world and allows us to make judgments. We know that a baby weighing 10.5 pounds is quite large, and that a human baby is unlikely to weigh 10.5 ounces or 10.5 kilograms. The context makes the number meaningful.

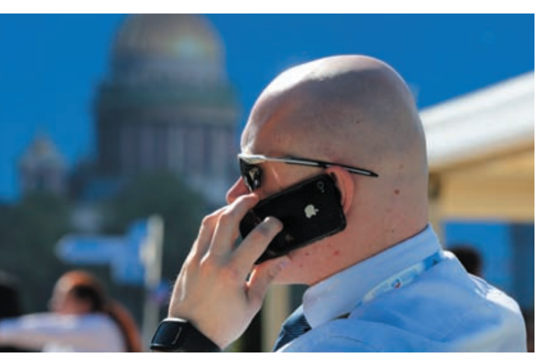
In your lifetime, you will be bombarded with data and statistical information. Poll results, television ratings, music sales, gas prices, unemployment rates, medical study outcomes, and standardized test scores are discussed daily in the media. Using data effectively is a large and growing part of most professions. A solid understanding of statistics will enable you to make sound, data-based decisions in your career and everyday life.

Data Beat Personal Experiences

It is tempting to base conclusions on your own experiences or the experiences of those you know. But our experiences may not be typical. In fact, the incidents that stick in our memory are often the unusual ones.

Do cell phones cause brain cancer?

Italian businessman Innocente Marcolini developed a brain tumor at age 60. He also talked on a cellular phone up to 6 hours per day for 12 years as part of his job. Mr. Marcolini's physician suggested that the brain tumor may have been caused by cell-phone use. So Mr. Marcolini decided to file suit in the Italian court system. A court ruled in his favor in October 2012.



Several statistical studies have investigated the link between cell-phone use and brain cancer. One of the largest was conducted by the Danish Cancer Society. Over 350,000 residents of Denmark were included in the study. Researchers compared the brain-cancer rate for the cell-phone users with the rate in the general population. The result: no statistical difference in brain-cancer rates.[1] In fact, most studies have produced similar conclusions. In spite of the evidence, many people (like Mr. Marcolini) are still convinced that cell phones can cause brain cancer.

In the public's mind, the compelling story wins every time. A statistically literate person knows better. ***Data are more reliable than personal experiences because they systematically describe an overall picture rather than focus on a few incidents.***

Where the Data Come from Matters

Are you kidding me?

The famous advice columnist Ann Landers once asked her readers, "If you had it to do over again, would you have children?" A few weeks later, her column was headlined "70% OF PARENTS SAY KIDS NOT WORTH IT." Indeed, 70% of the nearly 10,000 parents who wrote in said they would not have children if they could make the choice again. Do you believe that 70% of all parents regret having children?

You shouldn't. The people who took the trouble to write Ann Landers are not representative of all parents. Their letters showed that many of them were angry with their children. All we know from these data is that there are some unhappy parents out there. A statistically designed poll, unlike Ann Landers's appeal, targets specific people chosen in a way that gives all parents the same chance to be asked. Such a poll showed that 91% of parents *would* have children again.

Where data come from matters a lot. If you are careless about how you get your data, you may announce 70% "No" when the truth is close to 90% "Yes."

Who talks more—women or men?

According to Louann Brizendine, author of *The Female Brain*, women say nearly three times as many words per day as men. Skeptical researchers devised a study to test this claim. They used electronic devices to record the talking patterns of 396 university students from Texas, Arizona, and Mexico. The device was programmed to record 30 seconds of sound every 12.5 minutes without the carrier's knowledge. What were the results?

According to a published report of the study in *Scientific American*, "Men showed a slightly wider variability in words uttered. . . . But in the end, the sexes came out just about even in the daily averages: women at 16,215 words and men at 15,669."[2] When asked where she got her figures, Brizendine admitted that she used unreliable sources.[3]

The most important information about any statistical study is how the data were produced. Only carefully designed studies produce results that can be trusted.

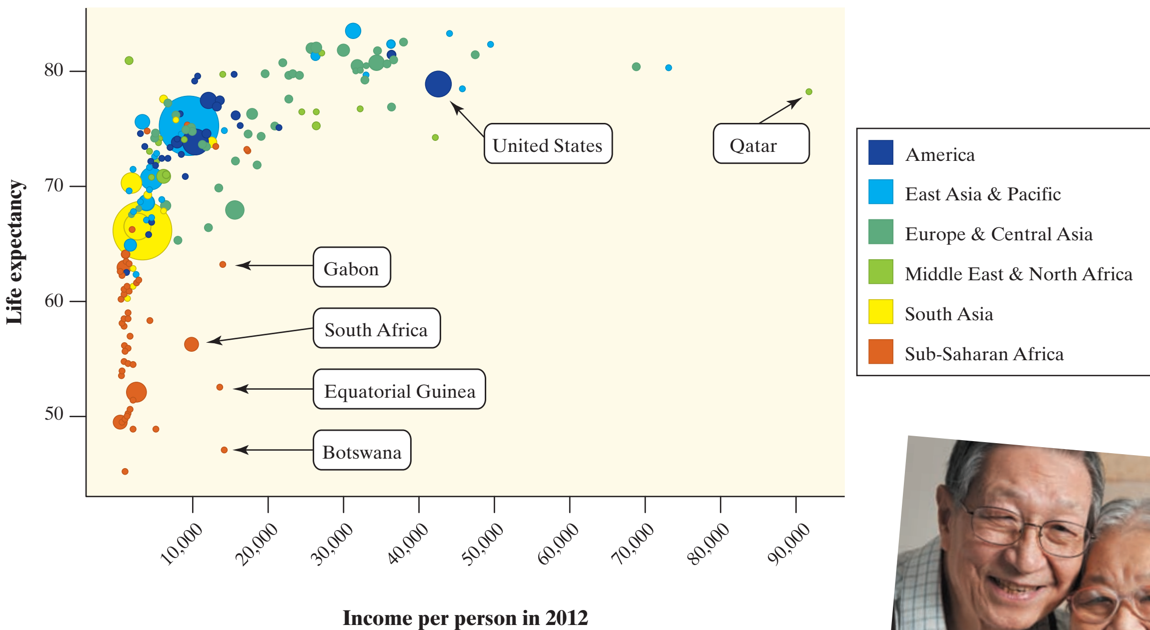
Always Plot Your Data

Yogi Berra, a famous New York Yankees baseball player known for his unusual quotes, had this to say: "You can observe a lot just by watching." That's a motto for learning from data. ***A carefully chosen graph is often more instructive than a bunch of numbers.***

Do people live longer in wealthier countries?

The Gapminder Web site, www.gapminder.org, provides loads of data on the health and well-being of the world's inhabitants. The graph on the next pages displays some data from Gapminder.[4] The individual points represent all the world's nations for which data are available. Each point shows the income per person and life expectancy in years for one country.

We expect people in richer countries to live longer. The overall pattern of the graph does show this, but the relationship has an interesting shape. Life expectancy rises very quickly as personal income increases and then levels off. People in very rich countries like the United States live no longer than people in poorer but not extremely poor nations. In some less wealthy countries, people live longer than in the United States. Several other nations stand out in the graph. What's special about each of these countries?



Graph of the life expectancy of people in many nations against each nation's income per person in 2012.

Variation Is Everywhere

Individuals vary. Repeated measurements on the same individual vary. Chance outcomes—like spins of a roulette wheel or tosses of a coin—vary. Almost everything varies over time. Statistics provides tools for understanding variation.

Have most students cheated on a test?

Researchers from the Josephson Institute were determined to find out. So they surveyed about 23,000 students from 100 randomly selected schools (both public and private) nationwide. The question they asked was "How many times have you cheated during a test at school in the past year?" Fifty-one percent said they had cheated at least once.[5]

If the researchers had asked the same question of *all* high school students, would exactly 51% have answered "Yes"? Probably not. If the Josephson Institute had selected a different sample of about 23,000 students to respond to the survey, they would probably have gotten a different estimate. *Variation is everywhere!*

Fortunately, statistics provides a description of how the sample results will vary in relation to the actual population percent. Based on the sampling method that this study used, we can say that the estimate of 51% is very likely to be within 1% of the true population value. That is, we can be quite confident that between 50% and 52% of *all* high school students would say that they have cheated on a test.

Because variation is everywhere, conclusions are uncertain. Statistics gives us a language for talking about uncertainty that is understood by statistically literate people everywhere.

Chapter 1

Exploring Data

case study

Do Pets or Friends Help Reduce Stress?

If you are a dog lover, having your dog with you may reduce your stress level. Does having a friend with you reduce stress? To examine the effect of pets and friends in stressful situations, researchers recruited 45 women who said they were dog lovers. Fifteen women were assigned at random to each of three groups: to do a stressful task alone, with a good friend present, or with their dogs present. The stressful task was to count backward by 13s or 17s. The woman's average heart rate during the task was one measure of the effect of stress. The table below shows the data.[1]

Average heart rates during stress with a pet (P), with a friend (F), and for the control group (C)

GROUP	RATE	GROUP	RATE	GROUP	RATE	GROUP	RATE
P	69.169	P	68.862	C	84.738	C	75.477
F	99.692	C	87.231	C	84.877	C	62.646
P	70.169	P	64.169	P	58.692	P	70.077
C	80.369	C	91.754	P	79.662	F	88.015
C	87.446	C	87.785	P	69.231	F	81.600
P	75.985	F	91.354	C	73.277	F	86.985
F	83.400	F	100.877	C	84.523	F	92.492
F	102.154	C	77.800	C	70.877	P	72.262
P	86.446	P	97.538	F	89.815	P	65.446
F	80.277	P	85.000	F	98.200		
C	90.015	F	101.062	F	76.908		
C	99.046	F	97.046	P	69.538		

Based on the data, does it appear that the presence of a pet or friend reduces heart rate during a stressful task? In this chapter, you'll develop the tools to help answer this question.

Introduction Data Analysis: Making Sense of Data

WHAT YOU WILL LEARN By the end of the section, you should be able to:

- Identify the individuals and variables in a set of data.
- Classify variables as categorical or quantitative.

Statistics is the science of data. The volume of data available to us is overwhelming. For example, the Census Bureau's American Community Survey collects data from 3,000,000 housing units each year. Astronomers work with data on tens of millions of galaxies. The checkout scanners at Walmart's 10,000 stores in 27 countries record hundreds of millions of transactions every week.

In all these cases, the data are trying to tell us a story—about U.S. households, objects in space, or Walmart shoppers. To hear what the data are saying, we need to help them speak by organizing, displaying, summarizing, and asking questions. That's **data analysis**.

Individuals and Variables

Any set of data contains information about some group of **individuals**. The characteristics we measure on each individual are called **variables**.

DEFINITION: Individuals and variables

Individuals are the objects described by a set of data. Individuals may be people, animals, or things.

A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

A high school's student data base, for example, includes data about every currently enrolled student. The students are the *individuals* described by the data set. For each individual, the data contain the values of *variables* such as age, gender, grade point average, homeroom, and grade level. In practice, any set of data is accompanied by background information that helps us understand the data. When you first meet a new data set, ask yourself the following questions:

1. *Who* are the individuals described by the data? How many individuals are there?
2. *What* are the variables? In what *units* are the variables recorded? Weights, for example, might be recorded in grams, pounds, thousands of pounds, or kilograms.

We could follow a newspaper reporter's lead and extend our list of questions to include *Why*, *When*, *Where*, and *How* were the data produced? For now, we'll focus on the first two questions.

Some variables, like gender and grade level, assign labels to individuals that place them into categories. Others, like age and grade point average (GPA), take numerical values for which we can do arithmetic. It makes sense to give an average GPA for a group of students, but it doesn't make sense to give an "average" gender.

DEFINITION: Categorical variable and quantitative variable

A **categorical variable** places an individual into one of several groups or categories.

A **quantitative variable** takes numerical values for which it makes sense to find an average.

AP® EXAM TIP If you learn to distinguish categorical from quantitative variables now, it will pay big rewards later. You will be expected to analyze categorical and quantitative variables correctly on the AP® exam.

Not every variable that takes number values is quantitative. Zip code is one example. Although zip codes are numbers, it doesn't make sense to talk about the average zip code. In fact, zip codes place individuals (people or dwellings) into categories based on location. Some variables—such as gender, race, and occupation—are categorical by nature. Other categorical variables are created by grouping values of a quantitative variable into classes. For instance, we could classify people in a data set by age: 0–9, 10–19, 20–29, and so on.

The proper method of analysis for a variable depends on whether it is categorical or quantitative. As a result, it is important to be able to distinguish these two types of variables. The type of data determines what kinds of graphs and which numerical summaries are appropriate.

EXAMPLE Census at School

Data, individuals, and variables

CensusAtSchool is an international project that collects data about primary and secondary school students using surveys. Hundreds of thousands of students from Australia, Canada, New Zealand, South Africa, and the United Kingdom have taken part in the project since 2000. Data from the surveys are available at the project's Web site (www.censusatschool.com). We used the site's "Random Data Selector" to choose 10 Canadian students who completed the survey in a recent year. The table below displays the data.

Province	Gender	Language spoken	Handed	Height (cm)	Wrist circum. (mm)	Preferred communication
Saskatchewan	Male	1	Right	175	180	In person
Ontario	Female	1	Right	162.5	160	In person
Alberta	Male	1	Right	178	174	Facebook
Ontario	Male	2	Right	169	160	Cell phone
Ontario	Female	2	Right	166	65	In person
Nunavut	Male	1	Right	168.5	160	Text messaging
Ontario	Female	1	Right	166	165	Cell phone
Ontario	Male	4	Left	157.5	147	Text Messaging
Ontario	Female	2	Right	150.5	187	Text Messaging
Ontario	Female	1	Right	171	180	Text Messaging

There is at least one suspicious value in the data table. We doubt that the girl who is 166 cm tall really has a wrist circumference of 65 mm (about 2.6 inches). Always look to be sure the values make sense!

PROBLEM:

(a) Who are the individuals in this data set?

(b) What variables were measured? Identify each as categorical or quantitative.

(c) Describe the individual in the highlighted row.

We'll see in Chapter 4 why choosing at random, as we did in this example, is a good idea.

SOLUTION:

(a) The individuals are the 10 randomly selected Canadian students who participated in the CensusAtSchool survey.

(b) The seven variables measured are the province where the student lives (categorical), gender (categorical), number of languages spoken (quantitative), dominant hand (categorical), height (quantitative), wrist circumference (quantitative), and preferred communication method (categorical).

(c) This student lives in Ontario, is male, speaks four languages, is left-handed, is 157.5 cm tall (about 62 inches), has a wrist circumference of 147 mm (about 5.8 inches), and prefers to communicate via text messaging.

For Practice *Try Exercise* 3

To make life simpler, we sometimes refer to "categorical data" or "quantitative data" instead of identifying the variable as categorical or quantitative.

Most data tables follow the format shown in the example—each row is an individual, and each column is a variable. Sometimes the individuals are called *cases*.

A variable generally takes values that vary (hence the name "variable"!). Categorical variables sometimes have similar counts in each category and sometimes don't. For instance, we might have expected similar numbers of males and females in the CensusAtSchool data set. But we aren't surprised to see that most students are right-handed. Quantitative variables may take values that are very close together or values that are quite spread out. We call the pattern of variation of a variable its **distribution**.

DEFINITION: Distribution

The **distribution** of a variable tells us what values the variable takes and how often it takes these values.

Section 1.1 begins by looking at how to describe the distribution of a single categorical variable and then examines relationships between categorical variables. Sections 1.2 and 1.3 and all of Chapter 2 focus on describing the distribution of a quantitative variable. Chapter 3 investigates relationships between two quantitative variables. In each case, we begin with graphical displays, then add numerical summaries for a more complete description.

HOW TO EXPLORE DATA

- Begin by examining each variable by itself. Then move on to study relationships among the variables.
- Start with a graph or graphs. Then add numerical summaries.

CHECK YOUR UNDERSTANDING

Jake is a car buff who wants to find out more about the vehicles that students at his school drive. He gets permission to go to the student parking lot and record some data. Later, he does some research about each model of car on the Internet. Finally, Jake makes a spreadsheet that includes each car's model, year, color, number of cylinders, gas mileage, weight, and whether it has a navigation system.

1. Who are the individuals in Jake's study?
2. What variables did Jake measure? Identify each as categorical or quantitative.

From Data Analysis to Inference

Sometimes, we're interested in drawing conclusions that go beyond the data at hand. That's the idea of **inference**. In the CensusAtSchool example, 9 of the 10 randomly selected Canadian students are right-handed. That's 90% of the *sample*. Can we conclude that 90% of the *population* of Canadian students who participated in CensusAtSchool are right-handed? No.

If another random sample of 10 students was selected, the percent who are right-handed might not be exactly 90%. Can we at least say that the actual population value is "close" to 90%? That depends on what we mean by "close."

The following Activity gives you an idea of how statistical inference works.

ACTIVITY | Hiring discrimination—it just won't fly!

MATERIALS:
Bag with 25 beads (15 of one color and 10 of another) or 25 identical slips of paper (15 labeled "M" and 10 labeled "F") for each student or pair of students

An airline has just finished training 25 pilots—15 male and 10 female—to become captains. Unfortunately, only eight captain positions are available right now. Airline managers announce that they will use a lottery to determine which pilots will fill the available positions. The names of all 25 pilots will be written on identical slips of paper. The slips will be placed in a hat, mixed thoroughly, and drawn out one at a time until all eight captains have been identified.

A day later, managers announce the results of the lottery. Of the 8 captains chosen, 5 are female and 3 are male. Some of the male pilots who weren't selected suspect that the lottery was not carried out fairly. One of these pilots asks your statistics class for advice about whether to file a grievance with the pilots' union.

The key question in this possible discrimination case seems to be: *Is it plausible (believable) that these results happened just by chance*? To find out, you and your classmates will *simulate* the lottery process that airline managers said they used.

1. Mix the beads/slips thoroughly. Without looking, remove 8 beads/slips from the bag. Count the number of female pilots selected. Then return the beads/slips to the bag.
2. Your teacher will draw and label a number line for a class *dotplot*. On the graph, plot the number of females you got in Step 1.
3. Repeat Steps 1 and 2 if needed to get a total of at least 40 simulated lottery results for your class.
4. Discuss the results with your classmates. Does it seem believable that airline managers carried out a fair lottery? What advice would you give the male pilot who contacted you?
5. Would your advice change if the lottery had chosen 6 female (and 2 male) pilots? What about 7 female pilots? Explain.

Our ability to do inference is determined by how the data are produced. Chapter 4 discusses the two main methods of data production—sampling and experiments—and the types of conclusions that can be drawn from each. As the Activity illustrates, the logic of inference rests on asking, "What are the chances?" *Probability*, the study of chance behavior, is the topic of Chapters 5 through 7. We'll introduce the most common inference techniques in Chapters 8 through 12.

Introduction Summary

- A data set contains information about a number of **individuals.** Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables.** A variable describes some characteristic of an individual, such as a person's height, gender, or salary.
- Some variables are **categorical** and others are **quantitative.** A categorical variable assigns a label that places each individual into one of several groups, such as male or female. A quantitative variable has numerical values that measure some characteristic of each individual, such as height in centimeters or salary in dollars.
- The **distribution** of a variable describes what values the variable takes and how often it takes them.

Introduction Exercises

The solutions to all exercises numbered in red are found in the Solutions Appendix, starting on page S-1.

1. **Protecting wood** How can we help wood surfaces resist weathering, especially when restoring historic wooden buildings? In a study of this question, researchers prepared wooden panels and then exposed them to the weather. Here are some of the variables recorded: type of wood (yellow poplar, pine, cedar); type of water repellent (solvent-based, water-based); paint thickness (millimeters); paint color (white, gray, light blue); weathering time (months). Identify each variable as categorical or quantitative.

2. **Medical study variables** Data from a medical study contain values of many variables for each of the people who were the subjects of the study. Here are some of the variables recorded: gender (female or male); age (years); race (Asian, black, white, or other); smoker (yes or no); systolic blood pressure (millimeters of mercury); level of calcium in the blood (micrograms per milliliter). Identify each as categorical or quantitative.

3. **A class survey** Here is a small part of the data set that describes the students in an AP® Statistics class. The data come from anonymous responses to a questionnaire filled out on the first day of class.

pg 3

Gender	Hand	Height (in.)	Homework time (min)	Favorite music	Pocket change (cents)
F	L	65	200	Hip-hop	50
M	L	72	30	Country	35
M	R	62	95	Rock	35
F	L	64	120	Alternative	0
M	R	63	220	Hip-hop	0
F	R	58	60	Alternative	76
F	R	67	150	Rock	215

(a) What individuals does this data set describe?

(b) What variables were measured? Identify each as categorical or quantitative.

(c) Describe the individual in the highlighted row.

4. **Coaster craze** Many people like to ride roller coasters. Amusement parks try to increase attendance by building exciting new coasters. The following table displays data on several roller coasters that were opened in a recent year.[2]

Roller coaster	Type	Height (ft)	Design	Speed (mph)	Duration (s)
Wild Mouse	Steel	49.3	Sit down	28	70
Terminator	Wood	95	Sit down	50.1	180
Manta	Steel	140	Flying	56	155
Prowler	Wood	102.3	Sit down	51.2	150
Diamondback	Steel	230	Sit down	80	180

(a) What individuals does this data set describe?

(b) What variables were measured? Identify each as categorical or quantitative.

(c) Describe the individual in the highlighted row.

5. **Ranking colleges** Popular magazines rank colleges and universities on their "academic quality" in serving undergraduate students. Describe two categorical variables and two quantitative variables that you might record for each institution.

6. **Students and TV** You are preparing to study the television-viewing habits of high school students. Describe two categorical variables and two quantitative variables that you might record for each student.

Multiple choice: Select the best answer.

Exercises 7 and 8 refer to the following setting. At the Census Bureau Web site www.census.gov, you can view detailed data collected by the American Community Survey. The following table includes data for 10 people chosen at random from the more than 1 million people in households contacted by the survey. "School" gives the highest level of education completed.

Weight (lb)	Age (yr)	Travel to work (min)	School	Gender	Income last year ($)
187	66	0	Ninth grade	1	24,000
158	66	n/a	High school grad	2	0
176	54	10	Assoc. degree	2	11,900
339	37	10	Assoc. degree	1	6000
91	27	10	Some college	2	30,000
155	18	n/a	High school grad	2	0
213	38	15	Master's degree	2	125,000
194	40	0	High school grad	1	800
221	18	20	High school grad	1	2500
193	11	n/a	Fifth grade	1	0

7. The individuals in this data set are

(a) households.

(b) people.

(c) adults.

(d) 120 variables.

(e) columns.

8. This data set contains

(a) 7 variables, 2 of which are categorical.

(b) 7 variables, 1 of which is categorical.

(c) 6 variables, 2 of which are categorical.

(d) 6 variables, 1 of which is categorical.

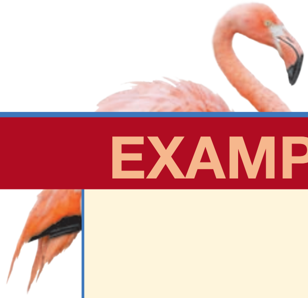
(e) None of these.

1.1 Analyzing Categorical Data

WHAT YOU WILL LEARN **By the end of the section, you should be able to:**

- Display categorical data with a bar graph. Decide if it would be appropriate to make a pie chart.
- Identify what makes some graphs of categorical data deceptive.
- Calculate and display the marginal distribution of a categorical variable from a two-way table.
- Calculate and display the conditional distribution of a categorical variable for a particular value of the other categorical variable in a two-way table.
- Describe the association between two categorical variables by comparing appropriate conditional distributions.

The values of a categorical variable are labels for the categories, such as "male" and "female." The distribution of a categorical variable lists the categories and gives either the *count* or the *percent* of individuals who fall within each category. Here's an example.



EXAMPLE

Radio Station Formats

Distribution of a categorical variable

The radio audience rating service Arbitron places U.S radio stations into categories that describe the kinds of programs they broadcast. Here are two different tables showing the distribution of station formats in a recent year:[3]

Frequency table

Format	Count of stations
Adult contemporary	1556
Adult standards	1196
Contemporary hit	569
Country	2066
News/Talk/Information	2179
Oldies	1060
Religious	2014
Rock	869
Spanish language	750
Other formats	1579
Total	**13,838**

Relative frequency table

Format	Percent of stations
Adult contemporary	11.2
Adult standards	8.6
Contemporary hit	4.1
Country	14.9
News/Talk/Information	15.7
Oldies	7.7
Religious	14.6
Rock	6.3
Spanish language	5.4
Other formats	11.4
Total	**99.9**

In this case, the *individuals* are the radio stations and the *variable* being measured is the kind of programming that each station broadcasts. The table on the left, which we call a **frequency table**, displays the counts (*frequencies*) of stations in each format category. On the right, we see a **relative frequency table** of the data that shows the percents (*relative frequencies*) of stations in each format category.

It's a good idea to check data for consistency. The counts should add to 13,838, the total number of stations. They do. The percents should add to 100%. In fact, they add to 99.9%. What happened? Each percent is rounded to the nearest tenth. The exact percents would add to 100, but the rounded percents only come close. This is **roundoff error.** Roundoff errors don't point to mistakes in our work, just to the effect of rounding off results.

Bar Graphs and Pie Charts

Columns of numbers take time to read. You can use a **pie chart** or a **bar graph** to display the distribution of a categorical variable more vividly. Figure 1.1 illustrates both displays for the distribution of radio stations by format.

Pie charts show the distribution of a categorical variable as a "pie" whose slices are sized by the counts or percents for the categories. A pie chart must include all the categories that make up a whole. In the radio station example, we needed the "Other formats" category to complete the whole (all radio stations) and allow us to make a pie chart. Use a pie chart only when you want to emphasize each

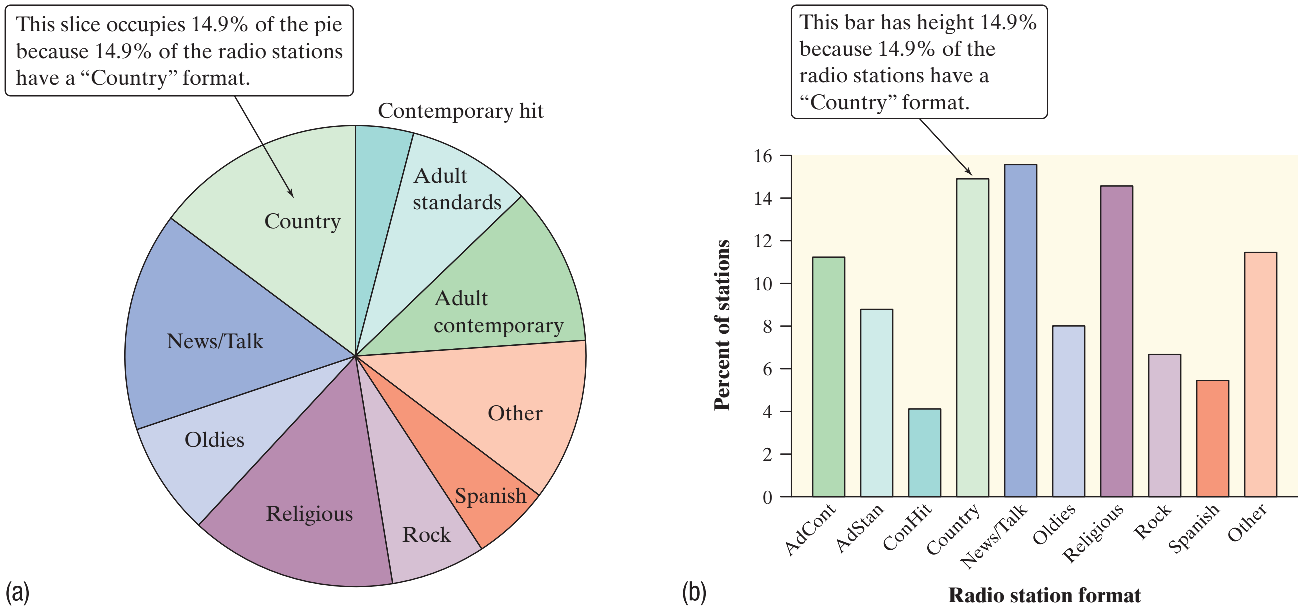
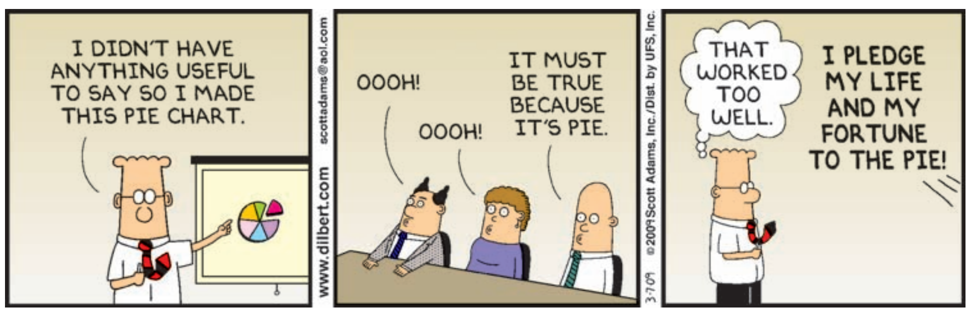


FIGURE 1.1 (a) Pie chart and (b) bar graph of U.S. radio stations by format.

category's relation to the whole. Pie charts are awkward to make by hand, but technology will do the job for you.



Bar graphs are also called *bar charts*.

Bar graphs represent each category as a bar. The bar heights show the category counts or percents. Bar graphs are easier to make than pie charts and are also easier to read. To convince yourself, try to use the pie chart in Figure 1.1 to estimate the percent of radio stations that have an "Oldies" format. Now look at the bar graph—it's easy to see that the answer is about 8%.

Bar graphs are also more flexible than pie charts. Both graphs can display the distribution of a categorical variable, but a bar graph can also compare any set of quantities that are measured in the same units.

EXAMPLE Who Owns an MP3 Player?

Choosing the best graph to display the data

Portable MP3 music players, such as the Apple iPod, are popular—but not equally popular with people of all ages. Here are the percents of people in various age groups who own a portable MP3 player, according to an Arbitron survey of 1112 randomly selected people.[4]