

Assumptions for Inference	And the Conditions That Support or Override Them
Proportions (z)	
<ul style="list-style-type: none"> • One sample <ol style="list-style-type: none"> 1. Individuals are independent. 2. Sample is sufficiently large. • Two groups <ol style="list-style-type: none"> 1. Groups are independent. 2. Data in each group are independent. 3. Both groups are sufficiently large. 	<ol style="list-style-type: none"> 1. SRS and $n < 10\%$ of the population. 2. Successes and failures each ≥ 10. 1. (Think about how the data were collected.) 2. Both are SRSs and $n < 10\%$ of populations OR random allocation. 3. Successes and failures each ≥ 10 for both groups.
Means (t)	
<ul style="list-style-type: none"> • One Sample ($df = n - 1$) <ol style="list-style-type: none"> 1. Individuals are independent. 2. Population has a Normal model. • Matched pairs ($df = n - 1$) <ol style="list-style-type: none"> 1. Data are matched. 2. Individuals are independent. 3. Population of differences is Normal. • Two independent groups (df from technology) <ol style="list-style-type: none"> 1. Groups are independent. 2. Data in each group are independent. 3. Both populations are Normal. 	<ol style="list-style-type: none"> 1. SRS and $n < 10\%$ of the population. 2. Histogram is unimodal and symmetric.* 1. (Think about the design.) 2. SRS and $n < 10\%$ OR random allocation. 3. Histogram of differences is unimodal and symmetric.* 1. (Think about the design.) 2. SRSs and $n < 10\%$ OR random allocation. 3. Both histograms are unimodal and symmetric.*
Distributions/Association (χ^2)	
<ul style="list-style-type: none"> • Goodness of fit ($df = \# \text{ of cells} - 1$; one variable, one sample compared with population model) <ol style="list-style-type: none"> 1. Data are counts. 2. Data in sample are independent. 3. Sample is sufficiently large. • Homogeneity [$df = (r - 1)(c - 1)$; many groups compared on one variable] <ol style="list-style-type: none"> 1. Data are counts. 2. Data in groups are independent. 3. Groups are sufficiently large. • Independence [$df = (r - 1)(c - 1)$; sample from one population classified on two variables] <ol style="list-style-type: none"> 1. Data are counts. 2. Data are independent. 3. Sample is sufficiently large. 	<ol style="list-style-type: none"> 1. (Are they?) 2. SRS and $n < 10\%$ of the population. 3. All expected counts ≥ 5. 1. (Are they?) 2. SRSs and $n < 10\%$ OR random allocation. 3. All expected counts ≥ 5. 1. (Are they?) 2. SRSs and $n < 10\%$ of the population. 3. All expected counts ≥ 5.
Regression (t, $df = n - 2$)	
<ul style="list-style-type: none"> • Association between two quantitative variables ($\beta = 0?$) <ol style="list-style-type: none"> 1. Form of relationship is linear. 2. Errors are independent. 3. Variability of errors is constant. 4. Errors have a Normal model. 	<ol style="list-style-type: none"> 1. Scatterplot looks approximately linear. 2. No apparent pattern in residuals plot. 3. Residuals plot has consistent spread. 4. Histogram of residuals is approximately unimodal and symmetric, or normal probability plot reasonably straight.*
(*less critical as n increases)	

Quick Guide to Inference

Think			Show				Tell?
Inference about?	One group or two?	Procedure	Model	Parameter	Estimate	SE	Chapter
Proportions	One sample	1-Proportion z-Interval	z	p	\hat{p}	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$	19
		1-Proportion z-Test				$\sqrt{\frac{p_0q_0}{n}}$	20, 21
	Two independent groups	2-Proportion z-Interval	z	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$	22
		2-Proportion z-Test				$\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}, \hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$	22
Means	One sample	t -Interval t -Test	t $df = n - 1$	μ	\bar{y}	$\frac{s}{\sqrt{n}}$	23
	Two independent groups	2-Sample t -Test 2-Sample t -Interval	t df from technology	$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	24
	Matched pairs	Paired t -Test Paired t -Interval	t $df = n - 1$	μ_d	\bar{d}	$\frac{s_d}{\sqrt{n}}$	25
Distributions <small>(one categorical variable)</small>	One sample	Goodness-of-Fit	χ^2 $df = cells - 1$	$\sum \frac{(Obs - Exp)^2}{Exp}$			26
	Many independent groups	Homogeneity χ^2 Test	χ^2 $df = (r - 1)(c - 1)$				
Independence <small>(two categorical variables)</small>	One sample	Independence χ^2 Test					
Association <small>(two quantitative variables)</small>	One sample	Linear Regression t -Test or Confidence Interval for β	t $df = n - 2$	β_1	b_1	$\frac{s_e}{s_x \sqrt{n - 1}}$ (compute with technology)	27
		*Confidence Interval for μ_ν		μ_ν	\hat{y}_ν	$\sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}$	
		*Prediction Interval for y_ν		y_ν	\hat{y}_ν	$\sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$	
Inference about?	One group or two?	Procedure	Model	Parameter	Estimate	SE	Chapter

Stats

Modeling the World

THIRD EDITION



EDITION

3

Stats

Modeling the World

David E. Bock

Ithaca High School
Cornell University

Paul F. Velleman

Cornell University

Richard D. De Veaux

Williams College

Addison-Wesley

Boston San Francisco New York

London Toronto Sydney Tokyo Singapore Madrid

Mexico City Munich Paris Cape Town Hong Kong Montreal

<i>Editor in Chief</i>	Deirdre Lynch
<i>Acquisitions Editor</i>	Christopher Cummings
<i>Senior Editor, AP and Electives</i>	Andrea Sheehan
<i>Assistant Editor</i>	Christina Lepre
<i>Editorial Assistant</i>	Dana Jones
<i>Senior Project Editor</i>	Chere Bemelmans
<i>Senior Managing Editor</i>	Karen Wernholm
<i>Senior Production Supervisor</i>	Sheila Spinney
<i>Cover Design</i>	Barbara T. Atkinson
<i>Digital Assets Manager</i>	Marianne Groth
<i>Media Producer</i>	Christine Stavrou
<i>Software Development</i>	Edward Chappell (MathXL) and Marty Wright (TestGen)
<i>Marketing Manager</i>	Alex Gay
<i>Marketing Coordinator</i>	Kathleen DeChavez
<i>Senior Author Support/Technology Specialist</i>	Joe Vetere
<i>Senior Prepress Supervisor</i>	Caroline Fell
<i>Senior Manufacturing Buyer</i>	Carol Melville
<i>Senior Media Buyer</i>	Ginny Michaud
<i>Production Coordination, Composition, and Illustrations</i>	Pre-Press PMG
<i>Interior Design</i>	The Davis Group, Inc.
<i>Cover Photo</i>	Pete McArthur

Library of Congress Cataloging-in-Publication Data

Bock, David E.

Stats : modeling the world / David E. Bock, Paul F. Velleman, Richard D. De Veaux.— 3rd ed.
p. cm.

Includes index.

ISBN 13: 978-0-13-135958-1

ISBN 10: 0-13-135958-4

1. Graphic calculators—Textbooks. I. Velleman, Paul F., 1949- II. De Veaux, Richard D. III. Title.

QA276.12.B628 2010

519.5—dc22

2008029019

For permission to use copyrighted material, grateful acknowledgement has been made to the copyright holders listed in Appendix D, which is hereby made part of this copyright page.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Addison-Wesley was aware of a trademark claim, the designations have been printed in initial caps or all caps. TI-Nspire and the TI-Nspire logo are trademarks of Texas Instruments, Inc.

Copyright © 2010, 2007, 2004 Pearson Education, Inc. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America. For information on obtaining permission for use of material in this work, please submit a written request to Pearson Education, Inc., Rights and Contracts Department, 501 Boylston Street, Boston, MA 02116, fax your request to 617-848-7047, or e-mail at <http://www.pearsoned.com/legal/permissions.htm>.

1 2 3 4 5 6 7 8 9 10—CRK—12 11 10 09

Addison-Wesley
is an imprint of



www.PearsonSchool.com/Advanced

ISBN 13: 978-0-13-135958-1

ISBN 10: 0-13-135958-4

*To Greg and Becca, great fun as kids and great friends as adults,
and especially to my wife and best friend, Joanna, for her
understanding, encouragement, and love*

—Dave

*To my sons, David and Zev, from whom I've learned so much,
and to my wife, Sue, for taking a chance on me*

—Paul

*To Sylvia, who has helped me in more ways than she'll ever know,
and to Nicholas, Scyrine, Frederick, and Alexandra,
who make me so proud in everything that they are and do*

—Dick

Meet the Authors



David E. Bock taught mathematics at Ithaca High School for 35 years. He has taught Statistics at Ithaca High School, Tompkins-Cortland Community College, Ithaca College, and Cornell University. Dave has won numerous teaching awards, including the MAA's Edyth May Sliffe Award for Distinguished High School Mathematics Teaching (twice), Cornell University's Outstanding Educator Award (three times), and has been a finalist for New York State Teacher of the Year.

Dave holds degrees from the University at Albany in Mathematics (B.A.) and Statistics/Education (M.S.). Dave has been a reader and table leader for the AP Statistics exam, serves as a Statistics consultant to the College Board, and leads workshops and institutes for AP Statistics teachers. He has recently served as K–12 Education and Outreach Coordinator and a senior lecturer for the Mathematics Department at Cornell University. His understanding of how students learn informs much of this book's approach.

Dave relaxes by biking and hiking. He and his wife have enjoyed many days camping across Canada and through the Rockies. They have a son, a daughter, and three grandchildren.



Paul F. Velleman has an international reputation for innovative Statistics education. He is the author and designer of the multimedia statistics CD-ROM *ActivStats*, for which he was awarded the EDUCOM Medal for innovative uses of computers in teaching statistics, and the ICTCM Award for Innovation in Using Technology in College Mathematics. He also developed the award-winning statistics program, Data Desk, and the Internet site Data And Story Library (DASL) (<http://dasl.datadesk.com>), which provides data sets for teaching Statistics. Paul's understanding of using and teaching with technology informs much of this book's approach.

Paul has taught Statistics at Cornell University since 1975. He holds an A.B. from Dartmouth College in Mathematics and Social Science, and M.S. and Ph.D. degrees in Statistics from Princeton University, where he studied with John Tukey. His research often deals with statistical graphics and data analysis methods. Paul co-authored (with David Hoaglin) *ABCs of Exploratory Data Analysis*. Paul is a Fellow of the American Statistical Association and of the American Association for the Advancement of Science.

Out of class, Paul sings baritone in a barbershop quartet. He is the father of two boys.



Richard D. De Veaux is an internationally known educator and consultant. He has taught at the Wharton School and the Princeton University School of Engineering, where he won a "Lifetime Award for Dedication and Excellence in Teaching." Since 1994, he has been Professor of Statistics at Williams College. Dick has won both the Wilcoxon and Shewell awards from the American Society for Quality. He is a fellow of the American Statistical Association. Dick is also well known in industry, where for the past 20 years he has consulted for such companies as Hewlett-Packard, Alcoa, DuPont, Pillsbury, General Electric, and Chemical Bank. He has also sometimes been called the "Official Statistician for the Grateful Dead." His real-world experiences and anecdotes illustrate many of this book's chapters.

Dick holds degrees from Princeton University in Civil Engineering (B.S.E.) and Mathematics (A.B.) and from Stanford University in Dance Education (M.A.) and Statistics (Ph.D.), where he studied with Persi Diaconis. His research focuses on the analysis of large data sets and data mining in science and industry.

In his spare time he is an avid cyclist and swimmer. He also is the founder and bass for the "Diminished Faculty," an a cappella Doo-Wop quartet at Williams College. Dick is the father of four children.

Contents

Preface ix



Exploring and Understanding Data 1

- CHAPTER 1 Stats Start Here 2
- CHAPTER 2 Data 7
- CHAPTER 3 Displaying and Describing Categorical Data 20
- CHAPTER 4 Displaying and Summarizing Quantitative Data 44
- CHAPTER 5 Understanding and Comparing Distributions 80
- CHAPTER 6 The Standard Deviation as a Ruler and the Normal Model 104
- Review of Part I Exploring and Understanding Data 135



Exploring Relationships Between Variables 145

- CHAPTER 7 Scatterplots, Association, and Correlation 146
- CHAPTER 8 Linear Regression 171
- CHAPTER 9 Regression Wisdom 201
- CHAPTER 10 Re-expressing Data: Get It Straight! 222
- Review of Part II Exploring Relationships Between Variables 244



Gathering Data 253

- CHAPTER 11 Understanding Randomness 255
- CHAPTER 12 Sample Surveys 268
- CHAPTER 13 Experiments and Observational Studies 292
- Review of Part III Gathering Data 317



Randomness and Probability 323

- CHAPTER 14 From Randomness to Probability 324
- CHAPTER 15 Probability Rules! 342
- CHAPTER 16 Random Variables 366
- CHAPTER 17 Probability Models 388
- Review of Part IV Randomness and Probability 405



From the Data at Hand to the World at Large 411

- CHAPTER 18 Sampling Distribution Models 412
- CHAPTER 19 Confidence Intervals for Proportions 439
- CHAPTER 20 Testing Hypotheses About Proportions 459
- CHAPTER 21 More About Tests and Intervals 480
- CHAPTER 22 Comparing Two Proportions 504
- Review of Part V From the Data at Hand to the World at Large 523



Learning About the World 529

- CHAPTER 23 Inferences About Means 530
- CHAPTER 24 Comparing Means 560
- CHAPTER 25 Paired Samples and Blocks 587
- Review of Part VI Learning About the World 609



Inference When Variables Are Related 617

- CHAPTER 26 Comparing Counts 618
- CHAPTER 27 Inferences for Regression 649
- Review of Part VII Inference When Variables Are Related 683
- CHAPTER 28 *Analysis of Variance—on the DVD
- CHAPTER 29 *Multiple Regression—on the DVD

Appendixes

- A Selected Formulas A-1
- B Guide to Statistical Software A-3
- C Answers A-25
- D Photo Acknowledgments A-59
- E Index A-61
- F TI Tips A-71
- G Tables A-73

*Indicates an optional chapter.

Preface

About the Book

We've been thrilled with the feedback we've received from teachers and students using *Stats: Modeling the World*, Second Edition. If there is a single hallmark of this book it is that students actually read it. We have reports from every level—from high school to graduate school—that students find our books easy and even enjoyable to read. We strive for a conversational, approachable style, and introduce anecdotes to maintain students' interest. And it works. Teachers report their amazement that students are voluntarily reading ahead of their assignments. Students write to tell us (to their amazement) that they actually enjoyed the book.

Stats: Modeling the World, Third Edition is written from the ground up with the understanding that Statistics is practiced with technology. This insight informs everything from our choice of forms for equations (favoring intuitive forms over calculation forms) to our extensive use of real data. Most important, it allows us to focus on teaching Statistical Thinking rather than calculation. The questions that motivate each of our hundreds of examples are not "how do you find the answer?" but "how do you think about the answer?"

Our Goal: Read This Book!

The best text in the world is of little value if students don't read it. Here are some of the ways we have made *Stats: Modeling the World*, Third Edition even more approachable:

- **Readability.** You'll see immediately that this book doesn't read like other Statistics texts. The style, both colloquial (with occasional humor) and informative, engages students to actually read the book to see what it says.
- **Informality.** Our informal diction doesn't mean that the subject matter is covered lightly or informally. We have tried to be precise and, wherever possible, to offer deeper explanations and justifications than those found in most introductory texts.
- **Focused lessons.** The chapters are shorter than in most other texts, to make it easier to focus on one topic at a time.
- **Consistency.** We've worked hard to avoid the "do what we say, not what we do" trap. From the very start we teach the importance of plotting data and checking assumptions and conditions, and we have been careful to model that behavior right through the rest of the book.
- **The need to read.** Students who plan just to skim the book may find our presentation a bit frustrating. The important concepts, definitions, and sample solutions don't sit in little boxes. This is a book that needs to be read, so we've tried to make the reading experience enjoyable.

New to the Third Edition

The third edition of *Stats: Modeling the World* continues and extends the successful innovations pioneered in our books, teaching Statistics and statistical thinking as it is practiced today. We've rewritten sections throughout the book to make them clearer and more interesting. We've introduced new up-to-the-minute motivating examples throughout. And, we've added a number of new features, each with the goal of making it even easier for students to put the concepts of Statistics together into a coherent whole.

FOR EXAMPLE

- ▶ **For Example.** In every chapter, you'll find approximately 4 new worked examples that illustrate how to apply new concepts and methods—**more than 100 new illustrative examples**. But these aren't isolated examples. We carry a discussion through the chapter with each *For Example*, picking up the story and moving it forward as students learn to apply each new concept.

STEP-BY-STEP EXAMPLE

- ▶ **Step-by-Step Worked Examples.** We've brought our innovative *Think/Show/Tell Step-by-Step* examples up-to-date with new examples and data.

A S

- ▶ **ActioStats Pointers.** In the third edition, the *ActioStats* pointers have been revised for clarity and now indicate exactly what they are pointing to—activity, video, simulation, or animation—paralleling the book's discussions to enhance learning.

TI-Nspire

- ▶ **TI-Nspire Activities.** We've created many demonstrations and investigations for TI-Nspire handhelds to enhance each chapter. They're on the DVD and at the book's Web site.

- ▶ **Exercises.** We've added **hundreds of new exercises**, including more single-concept exercises at the beginning of each set so students can be sure they have a clear understanding of each important topic before they're asked to tie them all together in more comprehensive exercises. Continuing exercises have been **updated with the most recent data**. Whenever possible, the data are on the DVD and the book's Web site so students can explore them further.

- ▶ **Data Sources.** Most of the data used in examples and exercises are from recent news stories, research articles, and other real-world sources. We've listed more of those sources in this edition.

- ▶ **Chapters 4 and 5** have been entirely rewritten and reorganized. We think you'll agree with our reviewers that the new organization—discussing displays and summaries for quantitative data in Chapter 4 and then expanding on those ideas to discuss comparisons across groups, outliers, and other more sophisticated topics in Chapter 5—provides a more exciting and interesting way to approach these fundamental topics.

- ▶ **Simulation.** We've improved the discussion of simulation in Chapter 11 so it could relate more easily to discussions of experimental design and probability. The simulations included in the *ActioStats* multimedia software on the book's DVD carry those ideas forward in a student-friendly fashion.

- ▶ **Teacher's Podcasts** (10 points in 10 minutes). Created and presented by the authors, these podcasts focus on key points in each chapter to help you with class preparation. These podcasts are available on the Instructor's Resource CD.

- ▶ **Video Lectures on DVD** featuring the textbook authors will help students review the high points of each chapter. Video presenters also work through examples from the text. The presentations feature the same student-friendly style and emphasis on critical thinking as the text.

Continuing Features



▶ **Think, Show, Tell.** The worked examples repeat the mantra of *Think, Show, and Tell* in every chapter. They emphasize the importance of thinking about a Statistics question (What do we know? What do we hope to learn? Are the assumptions and conditions satisfied?) and reporting our findings (the *Tell* step). The *Show* step contains the mechanics of calculating results and conveys our belief that it is only one part of the process. This rubric is highlighted in the *Step-by-Step* examples that guide the students through the process of analyzing the problem with the general explanation on the left and the worked-out problem on the right. The result is a better understanding of the concept, not just number crunching.



JUST CHECKING

▶ **Just Checking.** Within each chapter, we ask students to pause and think about what they've just read. These questions are designed to be a quick check that they understand the material. Answers are at the end of the exercise sets in each chapter so students can easily check themselves.



▶ **TI Tips.** We emphasize sound understanding of formulas and methods, but want students to use technology for actual calculations. Easy-to-read "TI Tips" in the chapters show students how to use TI-83/84 Plus statistics functions. (Help using a TI-89 or TI-Nspire appears in Appendix B.) We do remind students that calculators are just for "Show"—they cannot Think about what to do nor Tell what it all means.



▶ **Math Boxes.** In many chapters we present the mathematical underpinnings of the statistical methods and concepts. By setting these proofs, derivations, and justifications apart from the narrative, we allow the student to continue to follow the logical development of the topic at hand, yet also refer to the underlying mathematics for greater depth.



▶ **What Can Go Wrong?** Each chapter still contains our innovative *What Can Go Wrong?* sections that highlight the most common errors people make and the misconceptions they have about Statistics. Our goals are to help students avoid these pitfalls, and to arm them with the tools to detect statistical errors and to debunk misuses of statistics, whether intentional or not. In this spirit, some of our exercises probe the understanding of such failures.



▶ **What Have We Learned?** These chapter-ending summaries are great study guides providing complete overviews that highlight the new concepts, define the new terms, and list the skills that the student should have acquired in the chapter.

▶ **Exercises.** Throughout, we've maintained the pairing of examples so that each odd-numbered exercise (with an answer in the back of the book) is followed by an even-numbered exercise on the same concept. Exercises are still ordered by level of difficulty.



▶ **Reality Check.** We regularly remind students that Statistics is about understanding the world with data. Results that make no sense are probably wrong, no matter how carefully we think we did the calculations. Mistakes are often easy to spot with a little thought, so we ask students to stop for a reality check before interpreting their result.



▶ **Notation Alert.** Throughout this book we emphasize the importance of clear communication, and proper notation is part of the vocabulary of Statistics. We've found that it helps students when we call attention to the letters and symbols statisticians use to mean very specific things.



► **Connections.** Each chapter has a *Connections* section to link key terms and concepts with previous discussions and to point out continuing themes, helping students fit newly learned concepts into a growing understanding of Statistics.

► **On the Computer.** In the real world, Statistics is practiced with computers. We prefer not to choose a particular Statistics program. Instead, at the end of each chapter, we summarize what students can find in the most common packages, often with an annotated example. Computer output appearing in the book and in exercises is often generic, resembling all of the common packages to some degree.

ON THE COMPUTER

Coverage

Textbooks are often defined more by what they choose not to cover than by what they do cover. We've been guided in the choice and order of topics by several fundamental principles. First, we have tried to ensure that each new topic fits into the growing structure of understanding that we hope students will build. Several topic orders can support this goal. We explain our reasons for the topic order of the chapters in the ancillary Printed Test Bank and Resource Guide.

GAISE Guidelines. We have worked to provide materials to help each class, in its own way, follow the guidelines of the GAISE (Guidelines for Assessment and Instruction in Statistics Education) project sponsored by the American Statistical Association. That report urges that Statistics education should

1. emphasize Statistical literacy and develop Statistical thinking,
2. use real data,
3. stress conceptual understanding rather than mere knowledge of procedures,
4. foster active learning,
5. use technology for developing concepts and analyzing data, and
6. make assessment a part of the learning process.

We also have been guided by the syllabus of the AP* Statistics course. We agree with the wisdom of those who designed that course in their selection of topics and their emphasis on Statistics as a practical discipline. *Stats: Modeling the World* provides complete discussions of all AP* topics and teaches students communication skills that lead to success on the AP* examination. A correlation of the text to the AP* Statistics course standards is available in the Printed Test Bank and Resource Guide, on the Instructor's Resource CD, and at www.phschool.com/advanced/correlations/statistics.html.

Mathematics

Mathematics traditionally appears in Statistics texts in several roles:

1. It can provide a concise, clear statement of important concepts.
2. It can describe calculations to be performed with data.
3. It can embody proofs of fundamental results.

Of these, we emphasize the first. Mathematics can make discussions of Statistics concepts, probability, and inference clear and concise. We have tried to be sensitive to those who are discouraged by equations by also providing verbal descriptions and numerical examples.

This book is not concerned with proving theorems about Statistics. Some of these theorems are quite interesting, and many are important. Often, though, their proofs are not enlightening to introductory Statistics students, and can distract the audience from the concepts we want them to understand. However, we have not shied

away from the mathematics where we believed that it helped clarify without intimidating. You will find some important proofs, derivations, and justifications in Math Boxes that accompany the development of many topics.

Nor do we concentrate on calculations. Although statistics calculations are generally straightforward, they are also usually tedious. And, more to the point, they are often unnecessary. Today, virtually all statistics are calculated with technology, so there is little need for students to work by hand. The equations we use have been selected for their focus on understanding concepts and methods.

Technology and Data

To experience the real world of Statistics, it's best to explore real data sets using modern technology.

- ▶ **Technology.** We assume that you are using some form of technology in your Statistics course. That could be a calculator, a spreadsheet, or a statistics package. Rather than adopt any particular software, we discuss generic computer output. “TI-Tips”—included in most chapters—show students how to use statistics features of the TI-83/84 Plus series. The Companion DVD, included in the Teacher’s Edition, may be purchased for students and includes *ActivStats* and the software package Data Desk. Also, in Appendix B, we offer general guidance (by chapter) to help students get started on five common software platforms (Excel, MINITAB, Data Desk, JMP, and SPSS), a TI-89 calculator, and a TI-Nspire.
- ▶ **Data.** Because we use technology for computing, we don’t limit ourselves to small, artificial data sets. In addition to including some small data sets, we have built examples and exercises on real data with a moderate number of cases—usually more than you would want to enter by hand into a program or calculator. These data are included on the DVD as well as on the book’s Web site, www.aw.com/bock.

ON THE DVD

The DVD holds a number of supporting materials, including *ActivStats*, the *Data Desk* statistics package, an Excel add-in (DDXL), all large data sets from the text formatted for the most popular technologies, and two additional chapters.

ActivStats (for Data Desk). The award-winning *ActivStats* multimedia program supports learning chapter by chapter. It complements the book with videos of real-word stories, worked examples, animated expositions of each of the major Statistics topics, and tools for performing simulations, visualizing inference, and learning to use statistics software. The new version of *ActivStats* includes

- improved navigation and a cleaner design that makes it easier to find and use tools such as the Index and Glossary
- more than **1000 homework exercises**, including many new exercises, plus answers to the “odd numbered” exercises. Many are from the text, providing the data already set up for calculations, and some are unique to *ActivStats*. Many exercises link to data files for each statistics package.
- **17 short video clips**, many new and updated
- **70 animated activities**
- **117 teaching applets**
- more than **300 data sets**

Supplements

STUDENT SUPPLEMENTS

The following supplements are available for purchase:

Graphing Calculator Manual, by Patricia Humphrey (Georgia Southern University) and John Diehl (Hinsdale Central High School), is organized to follow the sequence of topics in the text, and is an easy to-follow, step-by-step guide on how to use the TI-83/84 Plus, TI-89, and TI-Nspire™ graphing calculators. It provides worked-out examples to help students fully understand and use the graphing calculator. (ISBN-13: 978-0-321-57094-9; ISBN-10: 0-321-57094-4)

Pearson Education AP* Test Prep Series: Statistics by Anne Carroll, Ruth Carver, Susan Peters, and Janice Ricks, is written specifically to complement *Stats: Modeling the World, Third Edition, AP* Edition*, and to help students prepare for the AP* Statistics exam. Students can review topics that are discussed in *Stats: Modeling the World, Third Edition AP* Edition*, and are likely to appear on the Advanced Placement Exam. The guide also contains test-taking strategies as well as practice tests. (ISBN 13: 978-0-13-135964-2; ISBN-10: 0-13-135964-9)

Statistics Study Card is a resource for students containing important formulas, definitions, and tables that correspond precisely to the De Veaux/Velleman/Bock Statistics series. This card can work as a reference for completing homework assignments or as an aid in studying. (ISBN-13: 978-0-321-46370-8; ISBN-10: 0-321-46370-6)

Graphing Calculator Tutorial for Statistics will guide students through the keystrokes needed to most efficiently use their graphing calculator. Although based on the TI-84 Plus Silver Edition, operating system 2.30, the keystrokes for this calculator are identical to those on the TI-84 Plus, and very similar to the TI-83 and TI-83 Plus. This tutorial should be helpful to students using any of these calculators, though there may be differences in some lessons. The tutorial is organized by topic. (ISBN-13: 978-0-321-41382-6; ISBN-10: 0-321-41382-2)

TEACHER SUPPLEMENTS

Most of the teacher supplements and resources for this book are available electronically. On adoption or to preview, please go to PearsonSchool.com/Advanced and click “Online Teacher Supplements.” You will be required to complete a one-time registration subject to verification before being emailed access information to download materials.

The following supplements are available to qualified adopters:

Teacher’s Edition contains answers to all exercises. Packaged with the Teacher’s Edition is the Companion DVD and the Instructor’s Resource CD. The Instructor’s Resource CD includes the Teachers’ Solutions Manual, Test Bank and Resource Guide, Audio Podcasts, PowerPoint slides, and Graphing Calculator Manual. (ISBN-13: 978-0-13-135959-8; ISBN-10: 0-13-135959-2)

Printed Test Bank and Resource Guide, by William Craine, contains chapter-by-chapter comments on the major concepts, tips on presenting topics (and what to avoid), teaching examples, suggested assignments, Web links and lists of other resources, as well as chapter quizzes, unit tests, investigative tasks, TI-Nspire activities, and suggestions for projects. An indispensable guide to help teachers prepare for class, the previous editions were soundly praised by new teachers of Statistics and seasoned veterans alike. The Printed Test Bank and Resource Guide is on the Instructor’s Resource CD and available for download. (ISBN-13: 978-0-13-135960-4; ISBN-10: 0-13-135960-6)

Teacher’s Solutions Manual, by William Craine, contains detailed solutions to all of the exercises. (ISBN-13: 978-0-13-136009-9; ISBN-10: 0-13-136009-4)

TestGen® CD enables teachers to build, edit, print, and administer tests using a computerized bank of questions developed to cover all the objectives of the text. TestGen is algorithmically based, allowing teachers to create multiple but equivalent versions of the same question or test with the click of a button. Teachers can also modify test bank questions or add new questions. Tests can be printed or administered online. (ISBN-13: 978-0-13-135961-1; ISBN-10: 0-13-135961-4)

PowerPoint Lecture Slides provide an outline to use in a lecture setting, presenting definitions, key concepts, and figures from the text. These slide are available on the Instructor’s Resource CD and available for download. (ISBN-13: 978-0-321-57101-4; ISBN 10: 0-321-57101-0)

Technology Resources

Instructor's Resource CD, packaged with every new Teacher's Edition, includes the Teacher's Solutions Manual, Test Bank and Resource Guide (which includes a correlation to the AP* Statistics course standards), Audio Podcasts, PowerPoint slides, and Graphing Calculator Manual. A replacement CD is available for purchase. (ISBN-13: 978-0-13-136349-6; ISBN-10: 0-13-136349-2)

Companion DVD A multimedia program on DVD designed to support learning chapter by chapter comes with the Teacher's Edition. It may be purchased separately for individual students or as a lab version (per work station). A replacement DVD is available for purchase. (ISBN-13: 978-0-13-136608-4; ISBN-10: 0-13-136608-4) The DVD holds a number of supporting materials, including:

- **ActivStats® for Data Desk.** The award-winning *ActivStats* multimedia program supports learning chapter by chapter with the book. It complements the book with videos of real-word stories, worked examples, animated expositions of each of the major Statistics topics, and tools for performing simulations, visualizing inference, and learning to use statistics software. The new version of *ActivStats* includes 17 short video clips; 170 animated activities and teaching applets; 300 data sets; 1,000 homework exercises, many with links to Data Desk files; interactive graphs, simulations, activities for the TI-Nspire graphing calculator, visualization tools, and much more.
- **Data Desk** statistics package.
- **TI-Nspire activities.** These investigations and demonstrations for the TI-Nspire handheld illustrate and explore important concepts from each chapter.
- **DDXL**, an Excel add-in, adds sound statistics and statistical graphics capabilities to Excel. DDXL adds, among other capabilities, boxplots, histograms, statistical scatterplots, normal probability plots, and statistical inference procedures not available in Excel's Data Analysis pack.
- **Data.** Data for exercises marked **T** are available on the DVD and at www.aw.com/bock formatted for Data Desk, Excel, JMP, MINITAB, SPSS, and the TI calculators, and as text files suitable for these and virtually any other statistics software.
- **Additional Chapters.** Two additional chapters cover **Analysis of Variance** (Chapter 28) and **Multiple Regression** (Chapter 29). These topics point the way to further study in Statistics.

ActivStats® The award-winning *ActivStats* multimedia program supports learning chapter by chapter with the book. It is available as a standalone DVD, or in a lab version (per work station). It complements the book with videos of real-word stories, worked examples, animated expositions of each of the major Statistics topics, and tools

for performing simulations, visualizing inference, and learning to use statistics software. The new version of *ActivStats* includes 17 short video clips; 170 animated activities and teaching applets; 300 data sets; 1,000 homework exercises, many with links to Data Desk files; interactive graphs, simulations, visualization tools, and much more. *ActivStats* (Mac and PC) is available in an all-in-one version for Data Desk, Excel, JMP, MINITAB, and SPSS. This DVD also includes Data Desk statistical software. For more information on options for purchasing *ActivStats*, contact Customer Service at 1-800-848-9500.

MathXL® for School is a powerful online homework, tutorial, and assessment system that accompanies Pearson textbooks in Statistics. With *MathXL for School*, teachers can create, edit, and assign online homework and tests using algorithmically generated exercises correlated at the objective level to the textbook. They can also create and assign their own online exercises and import TestGen tests for added flexibility. All student work is tracked in *MathXL for School's* online gradebook. Students can take chapter tests in *MathXL for School* and receive personalized study plans based on their test results. The study plan diagnoses weaknesses and links students directly to tutorial exercises for the objectives they need to study and retest. Students can also access supplemental animations directly from selected exercises. *MathXL for School* is available to qualified adopters. For more information, visit our Web site at www.MathXLforSchool.com, or contact your Pearson sales representative.

StatCrunch is a powerful online tool that provides an interactive environment for doing Statistics. *StatCrunch* can be used for both numerical and graphical data analysis, and uses interactive graphics to illustrate the connection between objects selected in a graph and the underlying data. *StatCrunch* may be purchased in a Registration Packet of 10 "redemptions." One redemption is for one student for 12 months beginning at the time of registration. Teacher access for *StatCrunch* adopters or for those wishing to preview the product may be obtained by filling out the form at www.pearsonschool.com/access_request (ISBN-13: 978-0-13-136416-5; ISBN-10: 0-13-136416-2)

Video Lectures on DVD with Subtitles feature the textbook authors reviewing the high points of each chapter. The presentations continue the same student-friendly style and emphasis on critical thinking as the text. The DVD format makes it easy and convenient to watch the videos from a computer at home or on campus. (ISBN 13: 978-0-321-57103-8; ISBN-10: 0-321-57103-7)

Companion Web Site (www.aw.com/bock) provides additional resources for instructors and students.

Acknowledgments

Many people have contributed to this book in all three of its editions. This edition would have never seen the light of day without the assistance of the incredible team at Addison-Wesley. Our editor in chief, Deirdre Lynch, was central to the genesis, development, and realization of the book from day one. Chris Cummings, acquisitions editor, provided much needed support. Chere Bemelmans, senior project editor, kept us on task as much as humanly possible. Sheila Spinney, senior production supervisor, kept the cogs from getting into the wheels where they often wanted to wander. Christina Lepre, assistant editor, and Kathleen DeChavez, marketing assistant, were essential in managing all of the behind-the-scenes work that needed to be done. Christine Stavrou, media producer, put together a top-notch media package for this book. Barbara T. Atkinson, senior designer, and Geri Davis are responsible for the wonderful way the book looks. Carol Melville, manufacturing buyer, and Ginny Michaud, senior media buyer, worked miracles to get this book and DVD in your hands, and Greg Tobin, publisher, was supportive and good-humored throughout all aspects of the project. Special thanks go out to Pre-Press PMG, the compositor, for the wonderful work they did on this book, and in particular to Laura Hakala, senior project manager, for her close attention to detail. We'd also like to thank our accuracy checkers whose monumental task was to make sure we said what we thought we were saying. They are Jackie Miller, The Ohio State University; Douglas Cashing, St. Bonaventure University; Jared Derksen, Rancho Cucamonga High School; and Susan Blackwell, First Flight High School.

We extend our sincere thanks for the suggestions and contributions made by the following reviewers of this edition:

Allen Back, *Cornell University, New York*

Susan Blackwell, *First Flight High School, North Carolina*

Kevin Crowther, *Lake Orion High School, Michigan*

Sam Erickson, *North High School, Wisconsin*

Guillermo Leon, *Coral Reef High School, Florida*

Martha Lowther, *The Tatnall School, Delaware*

Karl Ronning, *Davis Senior High School, California*

Agatha Shaw, *Valencia Community College, Florida*

We extend our sincere thanks for the suggestions and contributions made by the following reviewers, focus group participants, and class-testers of the previous edition:

John Arko, *Glenbrook South High School, IL*

Kathleen Arthur, *Shaker High School, NY*

Beverly Beemer, *Ruben S. Ayala High School, CA*

Judy Bevington, *Santa Maria High School, CA*

Susan Blackwell, *First Flight High School, NC*

Gail Brooks, *McLennan Community College, TX*

Walter Brown, *Brackenridge High School, TX*

Darin Clifft, *Memphis University School, TN*

Bill Craine, *Ithaca High School, NY*

Sybil Coley, *Woodward Academy, GA*

Caroline DiTullio, *Summit High School, NJ*

Jared Derksen, *Rancho Cucamonga High School, CA*

Laura Estersohn, *Scarsdale High School, NY*

Laura Favata, *Niskayuna High School, NY*

David Ferris, *Noblesville High School, IN*

Linda Gann, *Sandra Day O'Connor High School, TX*

Randall Groth, *Illinois State University, IL*

Donnie Hallstone, *Green River Community College, WA*

Howard W. Hand, *St. Marks School of Texas, TX*

Bill Hayes, *Foothill High School, CA*

Miles Hercamp, *New Palestine High School, IN*

Michelle Hipke, *Glen Burnie Senior High School, MD*

Carol Huss, *Independence High School, NC*

Sam Jovell, *Niskayuna High School, NY*

Peter Kaczmar, *Lower Merion High School, PA*

John Kotmel, *Lansing High School, NY*

Beth Lazerick, *St. Andrews School, FL*

Michael Legacy, *Greenhill School, TX*

John Lieb, *The Roxbury Latin School, MA*

John Maceli, *Ithaca College, NY*

Jim Miller, *Alta High School, UT*

Timothy E. Mitchell, *King Philip Regional High School, MA*

Maxine Nesbitt, *Carmel High School, IN*

Elizabeth Ann Przybysz, *Dr. Phillips High School, FL*

Diana Podhrasky, *Hillcrest High School, TX*

Rochelle Robert, *Nassau Community College, NY*

Bruce Saathoff, *Centennial High School, CA*

Murray Siegel, *Sam Houston State University, TX*

Chris Sollars, *Alamo Heights High School, TX*

Darren Starnes, *The Webb Schools, CA*

PART

I

Exploring and Understanding Data

Chapter 1

Stats Starts Here

Chapter 2

Data

Chapter 3

Displaying and Describing Categorical Data

Chapter 4

Displaying and Summarizing
Quantitative Data

Chapter 5

Understanding and Comparing
Distributions

Chapter 6

The Standard Deviation
as a Ruler and the
Normal Model

CHAPTER

1

Stats Starts Here¹



*“But where shall I begin?”
asked Alice. “Begin at the
beginning,” the King said
gravely, “and go on till you
come to the end: then stop.”*

—Lewis Carroll,
*Alice’s Adventures
in Wonderland*

Statistics gets no respect. People say things like “You can prove anything with Statistics.” People will write off a claim based on data as “just a statistical trick.” And Statistics courses don’t have the reputation of being students’ first choice for a fun elective.

But Statistics *is* fun. That’s probably not what you heard on the street, but it’s true. Statistics is about how to think clearly with data. A little practice thinking statistically is all it takes to start seeing the world more clearly and accurately.

So, What Is (Are?) Statistics?

Q: What is Statistics?

A: Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world.

Q: What are statistics?

A: Statistics (plural) are particular calculations made from data.

Q: So what is data?

A: You mean, “what *are* data?” Data is the plural form. The singular is datum.

Q: OK, OK, so what are data?

A: Data are values along with their context.

It seems every time we turn around, someone is collecting data on us, from every purchase we make in the grocery store, to every click of our mouse as we surf the Web. The United Parcel Service (UPS) tracks every package it ships from one place to another around the world and stores these records in a giant database. You can access part of it if you send or receive a UPS package. The database is about 17 terabytes big—about the same size as a database that contained every book in the Library of Congress would be. (But, we suspect, not *quite* as interesting.) What can anyone hope to do with all these data?

Statistics plays a role in making sense of the complex world in which we live today. Statisticians assess the risk of genetically engineered foods or of a new drug being considered by the Food and Drug Administration (FDA). They predict the number of new cases of AIDS by regions of the country or the number of customers likely to respond to a sale at the mall. And statisticians help scientists and social scientists understand how unemployment is related to environmental controls, whether enriched early education af-

¹ This chapter might have been called “Introduction,” but nobody reads the introduction, and we wanted you to read this. We feel safe admitting this here, in the footnote, because nobody reads footnotes either.

The ads say, "Don't drink and drive; you don't want to be a statistic." But you can't be a statistic.
 We say: "Don't be a datum."

facts later performance of school children, and whether vitamin C really prevents illness. Whenever there are data and a need for understanding the world, you need Statistics.

So our objectives in this book are to help you develop the insights to think clearly about the questions, use the tools to show what the data are saying, and acquire the skills to tell clearly what it all means.



FRAZZ reprinted by permission of United Feature Syndicate, Inc.

Statistics in a Word

Statistics is about variation.
 Data vary because we don't see everything and because even what we do see and measure, we measure imperfectly.
 So, in a very basic way, Statistics is about the real, imperfect world in which we live.

It can be fun, and sometimes useful, to summarize a discipline in only a few words. So,

- Economics is about . . . *Money (and why it is good).*
- Psychology: *Why we think what we think (we think).*
- Biology: *Life.*
- Anthropology: *Who?*
- History: *What, where, and when?*
- Philosophy: *Why?*
- Engineering: *How?*
- Accounting: *How much?*

In such a caricature, Statistics is about . . . **Variation.**

Data vary. People are different. We can't see everything, let alone measure it all. And even what we do measure, we measure imperfectly. So the data we wind up looking at and basing our decisions on provide, at best, an imperfect picture of the world. This fact lies at the heart of what Statistics is all about. How to make sense of it is a central challenge of Statistics.

So, How Will This Book Help?

A fair question. Most likely, this book will not turn out to be quite what you expected.

What's different?

Close your eyes and open the book to a page at random. Is there a graph or table on that page? Do that again, say, 10 times. We'll bet you saw data displayed in many ways, even near the back of the book and in the exercises.

We can better understand everything we do with data by making pictures. This book leads you through the entire process of thinking about a problem, finding and showing results, and telling others about what you have discovered. At each of these steps, we display data for better understanding and insight.

You looked at only a few randomly selected pages to get an impression of the entire book. We'll see soon that doing so was sound Statistics practice and reasoning.

Next, pick a chapter and read the first two sentences. (Go ahead; we'll wait.)

We'll bet you didn't see anything about Statistics. Why? Because the best way to understand Statistics is to see it at work. In this book, chapters usually start by presenting a story and posing questions. That's when Statistics really gets down to work.

There are three simple steps to doing Statistics right: *think, show, and tell*:



Think first. Know where you're headed and why. It will save you a lot of work.

Show is what most folks think Statistics is about. The *mechanics* of calculating statistics and making displays is important, but not the most important part of Statistics.

Tell what you've learned. Until you've explained your results so that someone else can understand your conclusions, the job is not done.

FOR EXAMPLE

STEP-BY-STEP

The best way to learn new skills is to take them out for a spin. In **For Example** boxes you'll see brief ways to apply new ideas and methods as you learn them. You'll also find more comprehensive worked examples called **Step-by-Steps**. These show you fully worked solutions side by side with commentary and discussion, modeling the way statisticians attack and solve problems. They illustrate how to think about the problem, what to show, and how to tell what it all means. These step-by-step examples will show you how to produce the kind of solutions instructors hope to see.

Sometimes, in the middle of the chapter, we've put a section called **Just Checking** . . . There you'll find a few short questions you can answer without much calculation—a quick way to check to see if you've understood the basic ideas in the chapter. You'll find the answers at the end of the chapter's exercises.



MATH BOX

Knowing where the formulas and procedures of Statistics come from and why they work will help you understand the important concepts. We'll provide brief, clear explanations of the mathematics that supports many of the statistical methods in **Math Boxes** like this.

TI Tips

How do I use
this thing?

Do statistics on your calculator!

Although we'll show you all the formulas you need to understand the calculations, you will most often use a calculator or computer to perform the mechanics of a statistics problem. Your graphing calculator has a specialized program called a "statistics package." Each chapter contains **TI Tips** that teach you how to use it (and avoid doing most of the messy calculations).

A S If you have the DVD, you'll find **ActivStats** parallels the chapters in this book and includes expanded lessons and activities to increase your understanding of the material covered in the text.

TI-Nspire

"Get your facts first, and then you can distort them as much as you please. (Facts are stubborn, but statistics are more pliable.)"

—Mark Twain



From time to time, you'll see an icon like this in the margin to signal that the *ActivStats* multimedia materials on the available DVD in the back of the book have an activity that you might find helpful at this point. Typically, we've flagged simulations and interactive activities because they're the most fun and will probably help you see how things work best. The chapters in *ActivStats* are the same as those in the text—just look for the named activity in the corresponding chapter.

If you are using TI-Nspire™ technology, these margin icons will alert you to activities and demonstrations that can help you understand important ideas in the text. If you have the DVD that's available with this book, you'll find these there; if not, they're also available on the book's Web site www.aw.com/bock.

One of the interesting challenges of Statistics is that, unlike in some math and science courses, there can be more than one right answer. This is why two statisticians can testify honestly on opposite sides of a court case. And it's why some people think that you can prove anything with statistics. But that's not true. People make mistakes using statistics, sometimes on purpose in order to mislead others. Most of the unintentional mistakes people make, though, are avoidable. We're not talking about arithmetic. More often, the mistakes come from using a method in the wrong situation or misinterpreting the results. Each chapter has a section called **What Can Go Wrong?** to help you avoid some of the most common mistakes.

Time out. From time to time, we'll take time out to discuss an interesting or important side issue. We indicate these by setting them apart like this.²

A S Introduction to (Your Statistics Package). *ActivStats* launches your statistics package (such as Data Desk) automatically. If you have the DVD, try it now.

ON THE COMPUTER

You'll find all sorts of stuff in margin notes, such as stories and quotations. For example:

"Computers are useless. They can only give you answers."

—Pablo Picasso

While Picasso underestimated the value of good statistics software, he did know that creating a solution requires more than just *Showing* an answer—it means you have to *Think* and *Tell*, too!

There are a number of statistics packages available for computers, and they differ widely in the details of how to use them and in how they present their results. But they all work from the same basic information and find the same results. Rather than adopt one package for this book, we present generic output and point out common features that you should look for. The . . . **on the Computer** section of most chapters (just before the exercises) holds this information. We also give a table of instructions to get you started on any of several commonly used packages, organized by chapters in Appendix B's Guide to Statistical Software.

At the end of each chapter, you'll see a brief summary of the important concepts you've covered in a section called **What Have We Learned?** That section includes a list of the **Terms** and a summary of the important **Skills** you've acquired in the chapter. You won't be able to learn the material from these summaries, but you can use them to check your knowledge of the important ideas in the chapter. If you have the skills, know the terms, and understand the concepts, you should be well prepared for the exam—and ready to use Statistics!

Beware: No one can learn Statistics just by reading or listening. The only way to learn it is to do it. So, of course, at the end of each chapter (except this one) you'll find **Exercises** designed to help you learn to use the Statistics you've just read about.

Some exercises are marked with an orange **T**. You'll find the data for these exercises on the DVD in the back of the book or on the book's Web site at www.aw.com/bock.

² Or in a footnote.

“Far too many scientists have only a shaky grasp of the statistical techniques they are using. They employ them as an amateur chef employs a cookbook, believing the recipes will work without understanding why. A more cordon bleu attitude . . . might lead to fewer statistical soufflés failing to rise.”

—*The Economist*, June 3, 2004, “**Sloppy stats shame science**”

We’ve paired up the exercises, putting similar ones together. So, if you’re having trouble doing an exercise, you will find a similar one either just before or just after it. You’ll find answers to the odd-numbered exercises at the back of the book. But these are only “answers” and not complete “solutions.” Huh? What’s the difference? The answers are sketches of the complete solutions. For most problems, your solution should follow the model of the Step-By-Step Examples. If your calculations match the numerical parts of the “answer” and your argument contains the elements shown in the answer, you’re on the right track. Your complete solution should explain the context, show your reasoning and calculations, and state your conclusions. Don’t fret too much if your numbers don’t match the printed answers to every decimal place. Statistics is more about getting the reasoning correct—pay more attention to how you interpret a result than what the digit in the third decimal place was.

In the real world, problems don’t come with chapters attached. So, in addition to the exercises at the ends of chapters, we’ve also collected a variety of problems at the end of each part of the text to make it more like the real world. This should help you to see whether you can sort out which methods to use when. If you can do that successfully, then you’ll know you understand Statistics.

Onward!

It’s only fair to warn you: You can’t get there by just picking out the highlighted sentences and the summaries. This book is different. It’s not about memorizing definitions and learning equations. It’s deeper than that. And much more fun. But . . .

*You have to read the book!*³

³ So, turn the page.



Many years ago, most stores in small towns knew their customers personally. If you walked into the hobby shop, the owner might tell you about a new bridge that had come in for your Lionel train set. The tailor knew your dad's size, and the hairdresser knew how your mom liked her hair. There are still some stores like that around today, but we're increasingly likely to shop at large stores, by phone, or on the Internet. Even so, when you phone an 800 number to buy new running shoes, customer service representatives may call you by your first name or ask about the socks you bought 6 weeks ago. Or the company may send an e-mail in October offering new head warmers for winter running. This company has millions of customers, and you called without identifying yourself. How did the sales rep know who you are, where you live, and what you had bought?

The answer is data. Collecting data on their customers, transactions, and sales lets companies track their inventory and helps them predict what their customers prefer. These data can help them predict what their customers may buy in the future so they know how much of each item to stock. The store can use the data and what it learns from the data to improve customer service, mimicking the kind of personal attention a shopper had 50 years ago.

Amazon.com opened for business in July 1995, billing itself as "Earth's Biggest Bookstore." By 1997, Amazon had a catalog of more than 2.5 million book titles and had sold books to more than 1.5 million customers in 150 countries. In 2006, the company's revenue reached \$10.7 billion. Amazon has expanded into selling a wide selection of merchandise, from \$400,000 necklaces¹ to yak cheese from Tibet to the largest book in the world.

Amazon is constantly monitoring and evolving its Web site to serve its customers better and maximize sales performance. To decide which changes to make to the site, the company experiments, collecting data and analyzing what works best. When you visit the Amazon Web site, you may encounter a different look or different suggestions and offers. Amazon statisticians want to know whether you'll follow the links offered, purchase the items suggested, or even spend a

"Data is king at Amazon. Clickstream and purchase data are the crown jewels at Amazon. They help us build features to personalize the Web site experience."

—Ronny Kohavi,
Director of Data Mining
and Personalization,
Amazon.com



¹ Please get credit card approval before purchasing online.

longer time browsing the site. As Ronny Kohavi, director of Data Mining and Personalization, said, “Data trumps intuition. Instead of using our intuition, we experiment on the live site and let our customers tell us what works for them.”

But What Are Data?

THE W’S:

WHO

WHAT

and in what units

WHEN

WHERE

WHY

HOW

We bet you thought you knew this instinctively. Think about it for a minute. What exactly *do* we mean by “data”?

Do data have to be numbers? The amount of your last purchase in dollars is numerical data, but some data record names or other labels. The names in Amazon.com’s database are data, but not numerical.

Sometimes, data can have values that look like numerical values but are just numerals serving as labels. This can be confusing. For example, the ASIN (Amazon Standard Item Number) of a book, like 0321570448, may have a numerical value, but it’s really just another name for *Stats: Modeling the World*.

Data values, no matter what kind, are useless without their context. Newspaper journalists know that the lead paragraph of a good story should establish the “Five W’s”: *Who, What, When, Where, and (if possible) Why*. Often we add *How* to the list as well. Answering these questions can provide the **context** for data values. The answers to the first two questions are essential. If you can’t answer *Who* and *What*, you don’t have **data**, and you don’t have any useful information.

Data Tables

Here are some data Amazon might collect:

B000001OAA	10.99	Chris G.	902	15783947	15.98	Kansas	Illinois	Boston
Canada	Samuel P.	Orange County	N	B000068ZVQ	Bad Blood	Nashville	Katherine H.	N
Mammals	10783489	Ohio	N	Chicago	12837593	11.99	Massachusetts	16.99
312	Monique D.	10675489	413	B0000015Y6	440	B000002BK9	Let Go	Y

A S **Activity: What Is (Are) Data?** Do you really know what’s data and what’s just numbers?

Try to guess what they represent. Why is that hard? Because these data have no *context*. If we don’t know *Who* they’re about or *What* they measure, these values are meaningless. We can make the meaning clear if we organize the values into a **data table** such as this one:

Purchase Order	Name	Ship to State/Country	Price	Area Code	Previous CD Purchase	Gift?	ASIN	Artist
10675489	Katharine H.	Ohio	10.99	440	Nashville	N	B0000015Y6	Kansas
10783489	Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
12837593	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
15783947	Monique D.	Canada	11.99	902	Let Go	N	B000001OAA	Mammals

Now we can see that these are four purchase records, relating to CD orders from Amazon. The column titles tell *What* has been recorded. The rows tell us *Who*. But be careful. Look at all the variables to see *Who* the variables are about. Even if people are involved, they may not be the *Who* of the data. For example, the *Who* here are the purchase orders (not the people who made the purchases).

A common place to find the *Who* of the table is the leftmost column. The other *W*'s might have to come from the company's database administrator.²

Who

In general, the rows of a data table correspond to individual **cases** about *Whom* (or about which—if they're not people) we record some characteristics. These cases go by different names, depending on the situation. Individuals who answer a survey are referred to as *respondents*. People on whom we experiment are *subjects* or (in an attempt to acknowledge the importance of their role in the experiment) *participants*, but animals, plants, Web sites, and other inanimate subjects are often just called *experimental units*. In a database, rows are called *records*—in this example, purchase records. Perhaps the most generic term is **cases**. In the Amazon table, the cases are the individual CD orders.

AS **Activity: Consider the Context** . . . Can you tell who's *Who* and what's *What*? And *Why*? This activity offers real-world examples to help you practice identifying the context.

Sometimes people just refer to data values as *observations*, without being clear about the *Who*. Be sure you know the *Who* of the data, or you may not know what the data say.

Often, the cases are a **sample** of cases selected from some larger **population** that we'd like to understand. Amazon certainly cares about its customers, but also wants to know how to attract all those other Internet users who may never have made a purchase from Amazon's site. To be able to generalize from the sample of cases to the larger population, we'll want the sample to be *representative* of that population—a kind of snapshot image of the larger world.

FOR EXAMPLE

Identifying the "Who"

In March 2007, *Consumer Reports* published an evaluation of large-screen, high-definition television sets (HDTVs). The magazine purchased and tested 98 different models from a variety of manufacturers.

Question: Describe the population of interest, the sample, and the *Who* of this study.

The magazine is interested in the performance of all HDTVs currently being offered for sale. It tested a sample of 98 sets, the "Who" for these data. Each HDTV set represents all similar sets offered by that manufacturer.

What and Why

The characteristics recorded about each individual are called **variables**. These are usually shown as the columns of a data table, and they should have a name that identifies *What* has been measured. Variables may seem simple, but to really understand your variables, you must *Think* about what you want to know.

Although area codes are numbers, do we use them that way? Is 610 twice 305? Of course it is, but is that the question? Why would we want to know whether Allentown, PA (area code 610), is twice Key West, FL (305)? Variables play different roles, and you can't tell a variable's role just by looking at it.

Some variables just tell us what group or category each individual belongs to. Are you male or female? Pierced or not? . . . What kinds of things can we learn about variables like these? A natural start is to *count* how many cases belong in each category. (Are you listening to music while reading this? We could count

²In database management, this kind of information is called "metadata."

It is wise to be careful. The *What* and *Why* of area codes are not as simple as they may first seem. When area codes were first introduced, AT&T was still the source of all telephone equipment, and phones had dials.



To reduce wear and tear on the dials, the area codes with the lowest digits (for which the dial would have to spin least) were assigned to the most populous regions—those with the most phone numbers and thus the area codes most likely to be dialed. New York City was assigned 212, Chicago 312, and Los Angeles 213, but rural upstate New York was given 607, Joliet was 815, and San Diego 619. For that reason, at one time the numerical value of an area code could be used to guess something about the population of its region. Now that phones have push-buttons, area codes have finally become just categories.

By international agreement, the International System of Units links together all systems of weights and measures. There are seven base units from which all other physical units are derived:

• Distance	Meter
• Mass	Kilogram
• Time	Second
• Electric current	Ampere
• Temperature	°Kelvin
• Amount of substance	Mole
• Intensity of light	Candela

AS **Activity: Recognize variables measured in a variety of ways.** This activity shows examples of the many ways to measure data.

AS **Activities: Variables.** Several activities show you how to begin working with data in your statistics package.

the number of students in the class who were and the number who weren't.) We'll look for ways to compare and contrast the sizes of such categories.

Some variables have measurement **units**. Units tell how each value has been measured. But, more importantly, units such as yen, cubits, carats, angstroms, nanoseconds, miles per hour, or degrees Celsius tell us the *scale* of measurement. The units tell us how much of something we have or how far apart two values are. Without units, the values of a measured variable have no meaning. It does little good to be promised a raise of 5000 a year if you don't know whether it will be paid in euros, dollars, yen, or Estonian krooni.

What kinds of things can we learn about measured variables? We can do a lot more than just counting categories. We can look for patterns and trends. (How much did you pay for your last movie ticket? What is the range of ticket prices available in your town? How has the price of a ticket changed over the past 20 years?)

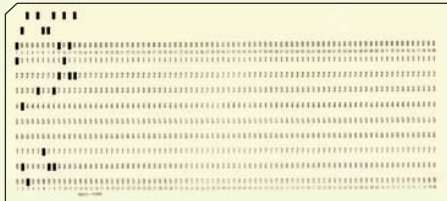
When a variable names categories and answers questions about how cases fall into those categories, we call it a **categorical variable**.³ When a measured variable with units answers questions about the quantity of what is measured, we call it a **quantitative variable**. These types can help us decide what to do with a variable, but they are really more about what we hope to learn from a variable than about the variable itself. It's the questions we ask a variable (the *Why* of our analysis) that shape how we think about it and how we treat it.

Some variables can answer questions only about categories. If the values of a variable are words rather than numbers, it's a good bet that it is categorical. But some variables can answer both kinds of questions. Amazon could ask for your *Age* in years. That seems quantitative, and would be if the company wanted to know the average age of those customers who visit their site after 3 a.m. But suppose Amazon wants to decide which CD to offer you in a special deal—one by Raffi, Blink-182, Carly Simon, or Mantovani—and needs to be sure to have adequate supplies on hand to meet the demand. Then thinking of your age in one of the categories—child, teen, adult, or senior—might be more useful. If it isn't clear whether a variable is categorical or quantitative, think about *Why* you are looking at it and what you want it to tell you.

A typical course evaluation survey asks, "How valuable do you think this course will be to you?": 1 = Worthless; 2 = Slightly; 3 = Middling; 4 = Reasonably; 5 = Invaluable. Is *Educational Value* categorical or quantitative? Once again, we'll look to the *Why*. A teacher might just count the number of students who gave each response for her course, treating *Educational Value* as a categorical variable. When she wants to see whether the course is improving, she might treat the responses as the *amount* of perceived value—in effect, treating the variable as quantitative. But what are the units? There is certainly an *order* of perceived worth: Higher numbers indicate higher perceived worth. A course that averages 4.5 seems more valuable than one that averages 2, but we should be careful about treating *Educational Value* as

³You may also see it called a *qualitative variable*.

One tradition that hangs on in some quarters is to name variables with cryptic abbreviations written in uppercase letters. This can be traced back to the 1960s, when the very first statistics computer programs were controlled with instructions punched on cards. The earliest punch card equipment used only uppercase letters, and the earliest statistics programs limited variable names to six or eight characters, so variables were called things like PRSRF3. Modern programs do not have such restrictive limits, so there is no reason for variable names that you wouldn't use in an ordinary sentence.



purely quantitative. To treat it as quantitative, she'll have to imagine that it has "educational value units" or some similar arbitrary construction. Because there are no natural units, she should be cautious. Variables like this that report order without natural units are often called "ordinal" variables. But saying "that's an ordinal variable" doesn't get you off the hook. You must still look to the *Why* of your study to decide whether to treat it as categorical or quantitative.

FOR EXAMPLE

Identifying "What" and "Why" of HDTVs.

Recap: A *Consumer Reports* article about 98 HDTVs lists each set's manufacturer, cost, screen size, type (LCD, plasma, or rear projection), and overall performance score (0–100).

Question: Are these variables categorical or quantitative? Include units where appropriate, and describe the "Why" of this investigation.

The "what" of this article includes the following variables:

- manufacturer (categorical);
- cost (in dollars, quantitative);
- screen size (in inches, quantitative);
- type (categorical);
- performance score (quantitative).

The magazine hopes to help consumers pick a good HDTV set.

Counts Count

In Statistics, we often count things. When Amazon considers a special offer of free shipping to customers, it might first analyze how purchases are shipped. They'd probably start by counting the number of purchases shipped by ground transportation, by second-day air, and by overnight air. Counting is a natural way to summarize the categorical variable *Shipping Method*. So every time we see counts, does that mean the variable is categorical? Actually, no.

We also use counts to measure the amounts of things. How many songs are on your digital music player? How many classes are you taking this semester? To measure these quantities, we'd naturally count. The variables (*Songs*, *Classes*) would be quantitative, and we'd consider the units to be "number of . . ." or, generically, just "counts" for short.

So we use counts in two different ways. When we count the cases in each category of a categorical variable, the category labels are the *What* and the individuals counted are the *Who* of our data. The counts themselves are not the

AS **Activity: Collect data in an experiment on yourself.** With the computer, you can experiment on yourself and then save the data. Go on to the subsequent related activities to check your understanding.

data, but are something we summarize about the data. Amazon counts the number of purchases in each category of the categorical variable *Shipping Method*. For this purpose (the *Why*), the *What* is shipping method and the *Who* is purchases.

Shipping Method	Number of Purchases
Ground	20,345
Second-day	7,890
Overnight	5,432

Other times our focus is on the amount of something, which we measure by counting. Amazon might record the number of teenage customers visiting their site each month to track customer growth and forecast CD sales (the *Why*). Now the *What* is *Teens*, the *Who* is *Months*, and the units are *Number of Teenage Customers*. *Teen* was a category when we looked at the categorical variable *Age*. But now it is a quantitative variable in its own right whose amount is measured by counting the number of customers.

Month	Number of Teenage Customers
January	123,456
February	234,567
March	345,678
April	456,789
May	...
...	...

Identifying Identifiers

What's your student ID number? It is numerical, but is it a quantitative variable? No, it doesn't have units. Is it categorical? Yes, but it is a special kind. Look at how many categories there are and at how many individuals are in each. There are as many categories as individuals and only one individual in each category. While it's easy to count the totals for each category, it's not very interesting. Amazon wants to know who you are when you sign in again and doesn't want to confuse you with some other customer. So it assigns you a unique identifier.

Identifier variables themselves don't tell us anything useful about the categories because we know there is exactly one individual in each. However, they are crucial in this age of large data sets. They make it possible to combine data from different sources, to protect confidentiality, and to provide unique labels. The variables *UPS Tracking Number*, *Social Security Number*, and Amazon's *ASIN* are all examples of identifier variables.

You'll want to recognize when a variable is playing the role of an identifier so you won't be tempted to analyze it. There's probably a list of unique ID numbers for students in a class (so they'll each get their own grade confidentially), but you might worry about the professor who keeps track of the average of these numbers from class to class. Even though this year's average ID number happens to be higher than last's, it doesn't mean that the students are better.

Where, When, and How

AS **Self-Test: Review concepts about data.** Like the Just Checking sections of this textbook, but interactive. (Usually, we won't reference the *ActivStats* self-tests here, but look for one whenever you'd like to check your understanding or review material.)

We must know *Who*, *What*, and *Why* to analyze data. Without knowing these three, we don't have enough to start. Of course, we'd always like to know more. The more we know about the data, the more we'll understand about the world.

If possible, we'd like to know the **When** and **Where** of data as well. Values recorded in 1803 may mean something different than similar values recorded last year. Values measured in Tanzania may differ in meaning from similar measurements made in Mexico.

How the data are collected can make the difference between insight and nonsense. As we'll see later, data that come from a voluntary survey on the Internet are almost always worthless. One primary concern of Statistics, to be discussed in Part III, is the design of sound methods for collecting data.

Throughout this book, whenever we introduce data, we'll provide a margin note listing the W's (and H) of the data. It's a habit we recommend. The first step of any data analysis is to know why you are examining the data (what you want to know), whom each row of your data table refers to, and what the variables (the columns of the table) record. These are the *Why*, the *Who*, and the *What*. Identifying them is a key part of the *Think* step of any analysis. Make sure you know all three before you proceed to *Show* or *Tell* anything about the data.



JUST CHECKING

In the 2003 Tour de France, Lance Armstrong averaged 40.94 kilometers per hour (km/h) for the entire course, making it the fastest Tour de France in its 100-year history. In 2004, he made history again by winning the race for an unprecedented sixth time. In 2005, he became the only 7-time winner and once again set a new record for the fastest average speed. You can find data on all the Tour de France races on the DVD. Here are the first three and last ten lines of the data set. Keep in mind that the entire data set has nearly 100 entries.

1. List as many of the W's as you can for this data set.
2. Classify each variable as categorical or quantitative; if quantitative, identify the units.



Year	Winner	Country of origin	Total time (h/min/s)	Avg. speed (km/h)	Stages	Total distance ridden (km)	Starting riders	Finishing riders
1903	Maurice Garin	France	94.33.00	25.3	6	2428	60	21
1904	Henri Cornet	France	96.05.00	24.3	6	2388	88	23
1905	Louis Trousselier	France	112.18.09	27.3	11	2975	60	24
:								
1999	Lance Armstrong	USA	91.32.16	40.30	20	3687	180	141
2000	Lance Armstrong	USA	92.33.08	39.56	21	3662	180	128
2001	Lance Armstrong	USA	86.17.28	40.02	20	3453	189	144
2002	Lance Armstrong	USA	82.05.12	39.93	20	3278	189	153
2003	Lance Armstrong	USA	83.41.12	40.94	20	3427	189	147
2004	Lance Armstrong	USA	83.36.02	40.53	20	3391	188	147
2005	Lance Armstrong	USA	86.15.02	41.65	21	3608	189	155
2006	Óscar Periero	Spain	89.40.27	40.78	20	3657	176	139
2007	Alberto Contador	Spain	91.00.26	38.97	20	3547	189	141
2008	Carlos Sastre	Spain	87.52.52	40.50	21	3559	199	145

There's a world of data on the Internet. These days, one of the richest sources of data is the Internet. With a bit of practice, you can learn to find data on almost any subject. Many of the data sets we use in this book were found in this way. The Internet has both advantages and disadvantages as a source of data. Among the advantages are the fact that often you'll be able to find even more current data than those we present. The disadvantage is that references to Internet addresses can "break" as sites evolve, move, and die.

Our solution to these challenges is to offer the best advice we can to help you search for the data, wherever they may be residing. We usually point you to a Web site. We'll sometimes suggest search terms and offer other guidance.

Some words of caution, though: Data found on Internet sites may not be formatted in the best way for use in statistics software. Although you may see a data table in standard form, an attempt to copy the data may leave you with a single column of values. You may have to work in your favorite statistics or spreadsheet program to reformat the data into variables. You will also probably want to remove commas from large numbers and such extra symbols as money indicators (\$, ¥, £); few statistics packages can handle these.

WHAT CAN GO WRONG?

- ▶ **Don't label a variable as categorical or quantitative without thinking about the question you want it to answer.** The same variable can sometimes take on different roles.
- ▶ **Just because your variable's values are numbers, don't assume that it's quantitative.** Categories are often given numerical labels. Don't let that fool you into thinking they have quantitative meaning. Look at the context.
- ▶ **Always be skeptical.** One reason to analyze data is to discover the truth. Even when you are told a context for the data, it may turn out that the truth is a bit (or even a lot) different. The context colors our interpretation of the data, so those who want to influence what you think may slant the context. A survey that seems to be about all students may in fact report just the opinions of those who visited a fan Web site. The question that respondents answered may have been posed in a way that influenced their responses.

TI Tips

Working with data

L1	L2	L3	1
71	-----	-----	
75			
75			
76			
80			
L1(6)=			

L1	L2	L3	1
71	-----	-----	
75			
75			
76			
80			

L1(4)=78			

You'll need to be able to enter and edit data in your calculator. Here's how.

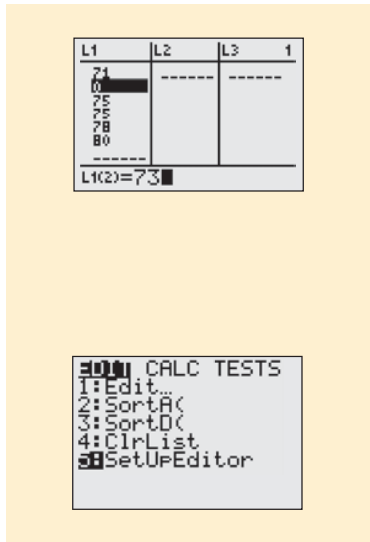
To enter data:

Hit the **STAT** button, and choose **EDIT** from the menu. You'll see a set of columns labeled **L1**, **L2**, and so on. Here is where you can enter, change, or delete a set of data.

Let's enter the heights (in inches) of the five starting players on a basketball team: 71, 75, 75, 76, and 80. Move the cursor to the space under **L1**, type in 71, and hit **ENTER** (or the down arrow). There's the first player. Now enter the data for the rest of the team.

To change a datum:

Suppose the 76" player grew since last season; his height should be listed as 78". Use the arrow keys to move the cursor onto the 76, then change the value and **ENTER** the correction.



To add more data:

We want to include the sixth man, 73" tall. It would be easy to simply add this new datum to the end of the list. However, sometimes the order of the data matters, so let's place this datum in numerical order. Move the cursor to the desired position (atop the first 75). Hit **2ND INS**, then **ENTER** the 73 in the new space.

To delete a datum:

The 78" player just quit the team. Move the cursor there. Hit **DEL**. Bye.

To clear the datalist:

Finished playing basketball? Move the cursor atop the **L1**. Hit **CLEAR**, then **ENTER** (or down arrow). You should now have a blank datalist, ready for you to enter your next set of values.

Lost a datalist?

Oops! Is **L1** now missing entirely? Did you delete **L1** by mistake, instead of just *clearing* it? Easy problem to fix: buy a new calculator. No? OK, then simply go to the **STAT EDIT** menu, and run **SetUpEditor** to recreate all the lists.



WHAT HAVE WE LEARNED?

We've learned that data are information in a context.

- ▶ The W's help nail down the context: *Who, What, Why, Where, When, and how*.
- ▶ We must know at least the *Who, What, and Why* to be able to say anything useful based on the data. The *Who* are the cases. The *What* are the *variables*. A variable gives information about each of the cases. The *Why* helps us decide which way to treat the variables.

We treat variables in two basic ways: as *categorical* or *quantitative*.

- ▶ Categorical variables identify a category for each case. Usually, we think about the counts of cases that fall into each category. (An exception is an identifier variable that just names each case.)
- ▶ Quantitative variables record measurements or amounts of something; they must have *units*.
- ▶ Sometimes we treat a variable as categorical or quantitative depending on what we want to learn from it, which means that some variables can't be pigeonholed as one type or the other. That's an early hint that in Statistics we can't always pin things down precisely.

Terms

Context	8. The context ideally tells <i>Who</i> was measured, <i>What</i> was measured, <i>How</i> the data were collected, <i>Where</i> the data were collected, and <i>When</i> and <i>Why</i> the study was performed.
Data	8. Systematically recorded information, whether numbers or labels, together with its context.
Data table	8. An arrangement of data in which each row represents a case and each column represents a variable.
Case	9. A case is an individual about whom or which we have data.
Population	9. All the cases we wish we knew about.
Sample	9. The cases we actually examine in seeking to understand the much larger population.
Variable	9. A variable holds information about the same characteristic for many cases.
Units	10. A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams.
Categorical variable	10. A variable that names categories (whether with words or numerals) is called categorical.
Quantitative variable	10. A variable in which the numbers act as numerical values is called quantitative. Quantitative variables always have units.

Skills

THINK

- ▶ Be able to identify the *Who*, *What*, *When*, *Where*, *Why*, and *How* of data, or recognize when some of this information has not been provided.
- ▶ Be able to identify the cases and variables in any data set.
- ▶ Be able to identify the population from which a sample was chosen.
- ▶ Be able to classify a variable as categorical or quantitative, depending on its use.
- ▶ For any quantitative variable, be able to identify the units in which the variable has been measured (or note that they have not been provided).

TELL

- ▶ Be able to describe a variable in terms of its *Who*, *What*, *When*, *Where*, *Why*, and *How* (and be prepared to remark when that information is not provided).

DATA ON THE COMPUTER

AS

Activity: Examine the Data. Take a look at your own data from your experiment (p. 12) and get comfortable with your statistics package as you find out about the experiment test results.

Most often we find statistics on a computer using a program, or *package*, designed for that purpose. There are many different statistics packages, but they all do essentially the same things. If you understand what the computer needs to know to do what you want and what it needs to show you in return, you can figure out the specific details of most packages pretty easily.

For example, to get your data into a computer statistics package, you need to tell the computer:

- ▶ Where to find the data. This usually means directing the computer to a file stored on your computer's disk or to data on a database. Or it might just mean that you have copied the data from a spreadsheet program or Internet site and it is currently on your computer's clipboard. Usually, the data should be in the form of a data table. Most computer statistics packages prefer the delimiter that marks the division between elements of a data table to be a tab character and the delimiter that marks the end of a case to be a return character.
- ▶ Where to put the data. (Usually this is handled automatically.)
- ▶ What to call the variables. Some data tables have variable names as the first row of the data, and often statistics packages can take the variable names from the first row automatically.

EXERCISES

1. **Voters.** A February 2007 Gallup Poll question asked, "In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?" The possible responses were "Democrat", "Republican", "Independent", "Other", and "No Response". What kind of variable is the response?
2. **Mood.** A January 2007 Gallup Poll question asked, "In general, do you think things have gotten better or gotten worse in this country in the last five years?" Possible answers were "Better", "Worse", "No Change", "Don't Know", and "No Response". What kind of variable is the response?
3. **Medicine.** A pharmaceutical company conducts an experiment in which a subject takes 100 mg of a substance orally. The researchers measure how many minutes it takes for half of the substance to exit the bloodstream. What kind of variable is the company studying?
4. **Stress.** A medical researcher measures the increase in heart rate of patients under a stress test. What kind of variable is the researcher studying?

(Exercises 5–12) For each description of data, identify *Who* and *What* were investigated and the *population of interest*.

Skills

THINK

- ▶ Be able to identify the *Who*, *What*, *When*, *Where*, *Why*, and *How* of data, or recognize when some of this information has not been provided.
- ▶ Be able to identify the cases and variables in any data set.
- ▶ Be able to identify the population from which a sample was chosen.
- ▶ Be able to classify a variable as categorical or quantitative, depending on its use.
- ▶ For any quantitative variable, be able to identify the units in which the variable has been measured (or note that they have not been provided).

TELL

- ▶ Be able to describe a variable in terms of its *Who*, *What*, *When*, *Where*, *Why*, and *How* (and be prepared to remark when that information is not provided).

DATA ON THE COMPUTER

AS

Activity: Examine the Data. Take a look at your own data from your experiment (p. 12) and get comfortable with your statistics package as you find out about the experiment test results.

Most often we find statistics on a computer using a program, or *package*, designed for that purpose. There are many different statistics packages, but they all do essentially the same things. If you understand what the computer needs to know to do what you want and what it needs to show you in return, you can figure out the specific details of most packages pretty easily.

For example, to get your data into a computer statistics package, you need to tell the computer:

- ▶ Where to find the data. This usually means directing the computer to a file stored on your computer's disk or to data on a database. Or it might just mean that you have copied the data from a spreadsheet program or Internet site and it is currently on your computer's clipboard. Usually, the data should be in the form of a data table. Most computer statistics packages prefer the delimiter that marks the division between elements of a data table to be a tab character and the delimiter that marks the end of a case to be a return character.
- ▶ Where to put the data. (Usually this is handled automatically.)
- ▶ What to call the variables. Some data tables have variable names as the first row of the data, and often statistics packages can take the variable names from the first row automatically.

EXERCISES

1. **Voters.** A February 2007 Gallup Poll question asked, "In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?" The possible responses were "Democrat", "Republican", "Independent", "Other", and "No Response". What kind of variable is the response?
 2. **Mood.** A January 2007 Gallup Poll question asked, "In general, do you think things have gotten better or gotten worse in this country in the last five years?" Possible answers were "Better", "Worse", "No Change", "Don't Know", and "No Response". What kind of variable is the response?
 3. **Medicine.** A pharmaceutical company conducts an experiment in which a subject takes 100 mg of a substance orally. The researchers measure how many minutes it takes for half of the substance to exit the bloodstream. What kind of variable is the company studying?
 4. **Stress.** A medical researcher measures the increase in heart rate of patients under a stress test. What kind of variable is the researcher studying?
- (Exercises 5–12) For each description of data, identify *Who* and *What* were investigated and the *population of interest*.

5. **The news.** Find a newspaper or magazine article in which some data are reported. For the data discussed in the article, answer the questions above. Include a copy of the article with your report.
6. **The Internet.** Find an Internet source that reports on a study and describes the data. Print out the description and answer the questions above.
7. **Bicycle safety.** Ian Walker, a psychologist at the University of Bath, wondered whether drivers treat bicycle riders differently when they wear helmets. He rigged his bicycle with an ultrasonic sensor that could measure how close each car was that passed him. He then rode on alternating days with and without a helmet. Out of 2500 cars passing him, he found that when he wore his helmet, motorists passed 3.35 inches closer to him, on average, than when his head was bare. [*NY Times*, Dec. 10, 2006]
8. **Investments.** Some companies offer 401(k) retirement plans to employees, permitting them to shift part of their before-tax salaries into investments such as mutual funds. Employers typically match 50% of the employees' contribution up to about 6% of salary. One company, concerned with what it believed was a low employee participation rate in its 401(k) plan, sampled 30 other companies with similar plans and asked for their 401(k) participation rates.
9. **Honesty.** Coffee stations in offices often just ask users to leave money in a tray to pay for their coffee, but many people cheat. Researchers at Newcastle University alternately taped two posters over the coffee station. During one week, it was a picture of flowers; during the other, it was a pair of staring eyes. They found that the average contribution was significantly higher when the eyes poster was up than when the flowers were there. Apparently, the mere feeling of being watched—even by eyes that were not real—was enough to encourage people to behave more honestly. [*NY Times*, Dec. 10, 2006]
10. **Movies.** Some motion pictures are profitable and others are not. Understandably, the movie industry would like to know what makes a movie successful. Data from 120 first-run movies released in 2005 suggest that longer movies actually make *less* profit.
11. **Fitness.** Are physically fit people less likely to die of cancer? An article in the May 2002 issue of *Medicine and Science in Sports and Exercise* reported results of a study that followed 25,892 men aged 30 to 87 for 10 years. The most physically fit men had a 55% lower risk of death from cancer than the least fit group.
12. **Molten iron.** The Cleveland Casting Plant is a large, highly automated producer of gray and nodular iron automotive castings for Ford Motor Company. The company is interested in keeping the pouring temperature of the molten iron (in degrees Fahrenheit) close to the specified value of 2550 degrees. Cleveland Casting measured the pouring temperature for 10 randomly selected crankshafts.

(Exercises 13–26) For each description of data, identify the *W*'s, name the variables, specify for each variable whether its use indicates that it should be treated as categorical or quantitative, and, for any quantitative variable, identify the units in which it was measured (or note that they were not provided).
13. **Weighing bears.** Because of the difficulty of weighing a bear in the woods, researchers caught and measured 54 bears, recording their weight, neck size, length, and sex. They hoped to find a way to estimate weight from the other, more easily determined quantities.
14. **Schools.** The State Education Department requires local school districts to keep these records on all students: age, race or ethnicity, days absent, current grade level, standardized test scores in reading and mathematics, and any disabilities or special educational needs.
15. **Arby's menu.** A listing posted by the Arby's restaurant chain gives, for each of the sandwiches it sells, the type of meat in the sandwich, the number of calories, and the serving size in ounces. The data might be used to assess the nutritional value of the different sandwiches.
16. **Age and party.** The Gallup Poll conducted a representative telephone survey of 1180 American voters during the first quarter of 2007. Among the reported results were the voter's region (Northeast, South, etc.), age, party affiliation, and whether or not the person had voted in the 2006 midterm congressional election.
17. **Babies.** Medical researchers at a large city hospital investigating the impact of prenatal care on newborn health collected data from 882 births during 1998–2000. They kept track of the mother's age, the number of weeks the pregnancy lasted, the type of birth (cesarean, induced, natural), the level of prenatal care the mother had (none, minimal, adequate), the birth weight and sex of the baby, and whether the baby exhibited health problems (none, minor, major).
18. **Flowers.** In a study appearing in the journal *Science*, a research team reports that plants in southern England are flowering earlier in the spring. Records of the first flowering dates for 385 species over a period of 47 years show that flowering has advanced an average of 15 days per decade, an indication of climate warming, according to the authors.
19. **Herbal medicine.** Scientists at a major pharmaceutical firm conducted an experiment to study the effectiveness of an herbal compound to treat the common cold. They exposed each patient to a cold virus, then gave them either the herbal compound or a sugar solution known to have no effect on colds. Several days later they assessed each patient's condition, using a cold severity scale ranging from 0 to 5. They found no evidence of the benefits of the compound.
20. **Vineyards.** Business analysts hoping to provide information helpful to American grape growers compiled these data about vineyards: size (acres), number of years in existence, state, varieties of grapes grown, average case price, gross sales, and percent profit.

- 21. **Streams.** In performing research for an ecology class, students at a college in upstate New York collect data on streams each year. They record a number of biological, chemical, and physical variables, including the stream name, the substrate of the stream (limestone, shale, or mixed), the acidity of the water (pH), the temperature (°C), and the BCI (a numerical measure of biological diversity).
- 22. **Fuel economy.** The Environmental Protection Agency (EPA) tracks fuel economy of automobiles based on information from the manufacturers (Ford, Toyota, etc.). Among the data the agency collects are the manufacturer, vehicle type (car, SUV, etc.), weight, horsepower, and gas mileage (mpg) for city and highway driving.
- 23. **Refrigerators.** In 2006, *Consumer Reports* published an article evaluating refrigerators. It listed 41 models, giving the brand, cost, size (cu ft), type (such as top freezer), estimated annual energy cost, an overall rating (good, excellent, etc.), and the repair history for that brand (percentage requiring repairs over the past 5 years).

- 24. **Walking in circles.** People who get lost in the desert, mountains, or woods often seem to wander in circles rather than walk in straight lines. To see whether people naturally walk in circles in the absence of visual clues, researcher Andrea Axtell tested 32 people on a football field. One at a time, they stood at the center of one goal line, were blindfolded, and then tried to walk to the other goal line. She recorded each individual’s sex, height, handedness, the number of yards each was able to walk before going out of bounds, and whether each wandered off course to the left or the right. No one made it all the way to the far end of the field without crossing one of the sidelines. [*STATS* No. 39, Winter 2004]

T 25. Horse race 2008. The Kentucky Derby is a horse race that has been run every year since 1875 at Churchill Downs, Louisville, Kentucky. The race started as a 1.5-mile race, but in 1896, it was shortened to 1.25 miles because experts felt that 3-year-old horses shouldn’t run such a long race that early in the season. (It has been run in May every year but one—1901—when it took place on April 29). Here are the data for the first four and several recent races.

Date	Winner	Margin (lengths)	Jockey	Winner’s Payoff (\$)	Duration (min:sec)	Track Condition
May 17, 1875	Aristides	2	O. Lewis	2850	2:37.75	Fast
May 15, 1876	Vagrant	2	B. Swim	2950	2:38.25	Fast
May 22, 1877	Baden-Baden	2	W. Walker	3300	2:38.00	Fast
May 21, 1878	Day Star	1	J. Carter	4050	2:37.25	Dusty
.....						
May 1, 2004	Smarty Jones	2 3/4	S. Elliott	854800	2:04.06	Sloppy
May 7, 2005	Giacomo	1/2	M. Smith	5854800	2:02.75	Fast
May 6, 2006	Barbaro	6 1/2	E. Prado	1453200	2:01.36	Fast
May 5, 2007	Street Sense	2 1/4	C. Borel	1450000	2:02.17	Fast
May 3, 2008	Big Brown	4 3/4	K. Desormeaux	1451800	2:01.82	Fast

- T 26. Indy 2008.** The 2.5-mile Indianapolis Motor Speedway has been the home to a race on Memorial Day nearly every year since 1911. Even during the first race, there were controversies. Ralph Mulford was given the checkered flag first but took three extra laps just to make sure he’d completed 500 miles. When he finished, another driver, Ray Harroun, was being presented with the

winner’s trophy, and Mulford’s protests were ignored. Harroun averaged 74.6 mph for the 500 miles. In 2008, the winner, Scott Dixon, averaged 143.567 mph.

Here are the data for the first five races and five recent Indianapolis 500 races. Included also are the pole winners (the winners of the trial races, when each driver drives alone to determine the position on race day).

Year	Winner	Pole Position	Average Speed (mph)	Pole Winner	Average Pole Speed (mph)
1911	Ray Harroun	28	74.602	Lewis Strang	.
1912	Joe Dawson	7	78.719	Gil Anderson	.
1913	Jules Goux	7	75.933	Caleb Bragg	.
1914	René Thomas	15	82.474	Jean Chassagne	.
1915	Ralph DePalma	2	89.840	Howard Wilcox	98.580
...					
2004	Buddy Rice	1	138.518	Buddy Rice	220.024
2005	Dan Wheldon	16	157.603	Tony Kanaan	224.308
2006	Sam Hornish Jr.	1	157.085	Sam Hornish Jr.	228.985
2007	Dario Franchitti	3	151.744	Hélio Castroneves	225.817
2008	Scott Dixon	1	143.567	Scott Dixon	221.514



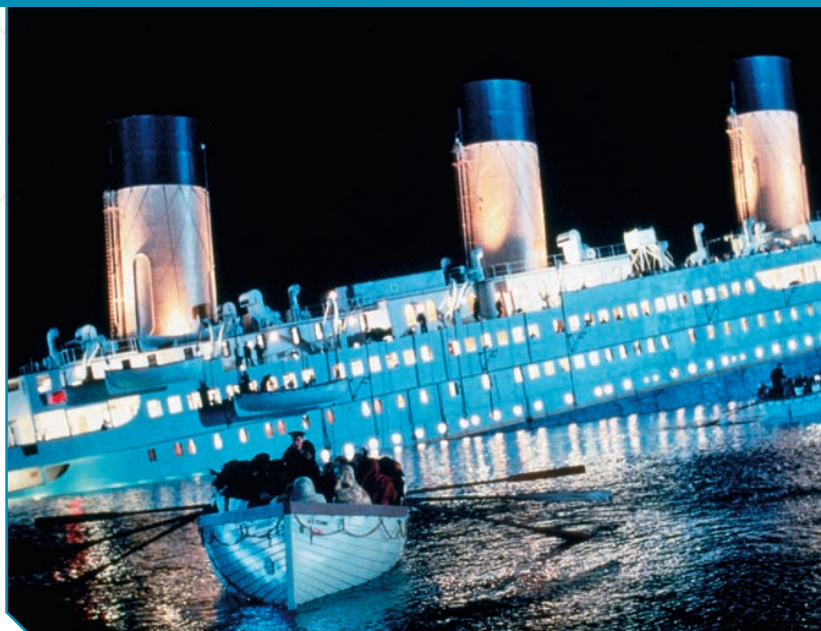
JUST CHECKING Answers

1. Who—Tour de France races; What—year, winner, country of origin, total time, average speed, stages, total distance ridden, starting riders, finishing riders; How—official statistics at race; Where—France (for the most part); When—1903 to 2008; Why—not specified (To see progress in speeds of cycling racing?)

2.

Variable	Type	Units
Year	Quantitative or Categorical	Years
Winner	Categorical	
Country of Origin	Categorical	
Total Time	Quantitative	Hours/minutes/seconds
Average Speed	Quantitative	Kilometers per hour
Stages	Quantitative	Counts (stages)
Total Distance	Quantitative	Kilometers
Starting Riders	Quantitative	Counts (riders)
Finishing Riders	Quantitative	Counts (riders)

Displaying and Describing Categorical Data



What happened on the *Titanic* at 11:40 on the night of April 14, 1912, is well known. Frederick Fleet’s cry of “Iceberg, right ahead” and the three accompanying pulls of the crow’s nest bell signaled the beginning of a nightmare that has become legend. By 2:15 a.m., the *Titanic*, thought by many to be unsinkable, had sunk, leaving more than 1500 passengers and crew members on board to meet their icy fate.

Here are some data about the passengers and crew aboard the *Titanic*. Each case (row) of the data table represents a person on board the ship. The variables are the person’s *Survival* status (Dead or Alive), *Age* (Adult or Child), *Sex* (Male or Female), and ticket *Class* (First, Second, Third, or Crew).

The problem with a data table like this—and in fact with all data tables—is that you can’t *see* what’s going on. And seeing is just what we want to do. We need ways to show the data so that we can see patterns, relationships, trends, and exceptions.

- WHO** People on the *Titanic*
- WHAT** Survival status, age, sex, ticket class
- WHEN** April 14, 1912
- WHERE** North Atlantic
- HOW** A variety of sources and Internet sites
- WHY** Historical interest

AS **Video: The Incident** tells the story of the *Titanic*, and includes rare film footage.

Survival	Age	Sex	Class
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Alive	Adult	Female	First
Dead	Adult	Male	Third
Dead	Adult	Male	Crew

Table 3.1

Part of a data table showing four variables for nine people aboard the *Titanic*.

The Three Rules of Data Analysis



FIGURE 3.1 A Picture to Tell a Story

Florence Nightingale (1820–1910), a founder of modern nursing, was also a pioneer in health management, statistics, and epidemiology. She was the first female member of the British Statistical Society and was granted honorary membership in the newly formed American Statistical Association.

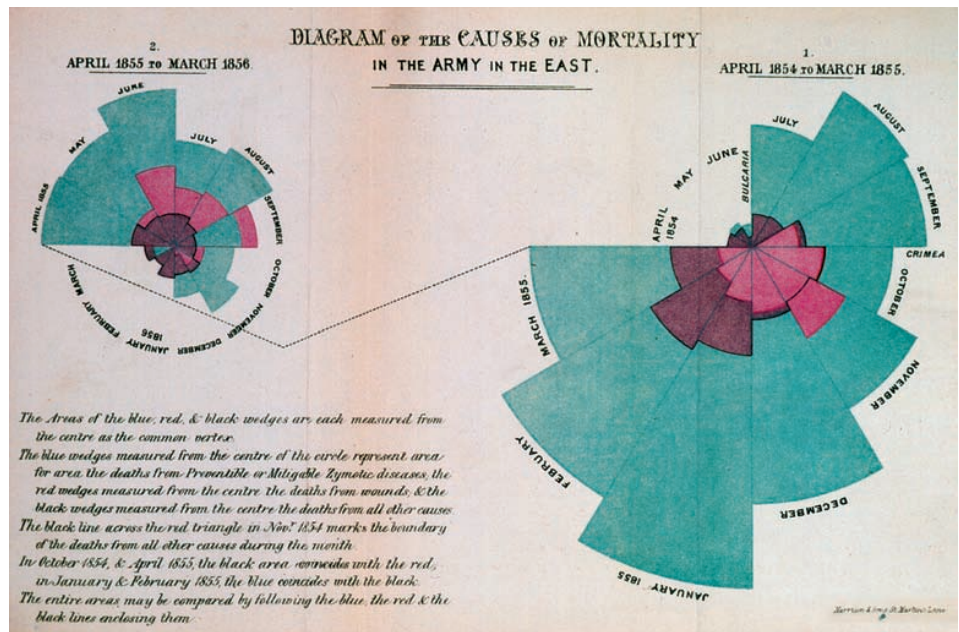
To argue forcefully for better hospital conditions for soldiers, she and her colleague, Dr. William Farr, invented this display, which showed that in the Crimean War, far more soldiers died of illness and infection than of battle wounds. Her campaign succeeded in improving hospital conditions and nursing for soldiers.

Florence Nightingale went on to apply statistical methods to a variety of important health issues and published more than 200 books, reports, and pamphlets during her long and illustrious career.

So, what should we do with data like these? There are three things you should always do first with data:

1. **Make a picture.** A display of your data will reveal things you are not likely to see in a table of numbers and will help you to *Think* clearly about the patterns and relationships that may be hiding in your data.
2. **Make a picture.** A well-designed display will *Show* the important features and patterns in your data. A picture will also show you the things you did not expect to see: the extraordinary (possibly wrong) data values or unexpected patterns.
3. **Make a picture.** The best way to *Tell* others about your data is with a well-chosen picture.

These are the three rules of data analysis. There are pictures of data throughout the book, and new kinds keep showing up. These days, technology makes drawing pictures of data easy, so there is no reason not to follow the three rules.



Frequency Tables: Making Piles

AS **Activity:** Make and examine a table of counts. Even data on something as simple as hair color can reveal surprises when you organize it in a data table.

Class	Count
First	325
Second	285
Third	706
Crew	885

Table 3.2

A frequency table of the *Titanic* passengers.

To make a picture of data, the first thing we have to do is to make piles. Making piles is the beginning of understanding about data. We pile together things that seem to go together, so we can see how the cases distribute across different categories. For categorical data, piling is easy. We just count the number of cases corresponding to each category and pile them up.

One way to put all 2201 people on the *Titanic* into piles is by ticket *Class*, counting up how many had each kind of ticket. We can organize these counts into a **frequency table**, which records the totals and the category names.

Even when we have thousands of cases, a variable like ticket *Class*, with only a few categories, has a frequency table that's easy to read. A frequency table with dozens or hundreds of categories would be much harder to read. We use the names of the categories to label each row in the frequency table. For ticket *Class*, these are "First," "Second," "Third," and "Crew."

Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

Table 3.3
A relative frequency table for the same data.

Counts are useful, but sometimes we want to know the fraction or **proportion** of the data in each category, so we divide the counts by the total number of cases. Usually we multiply by 100 to express these proportions as **percentages**. A **relative frequency table** displays the *percentages*, rather than the counts, of the values in each category. Both types of tables show how the cases are distributed across the categories. In this way, they describe the **distribution** of a categorical variable because they name the possible categories and tell how frequently each occurs.

The Area Principle

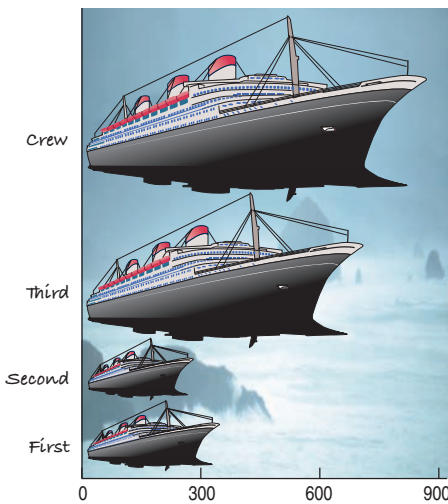


FIGURE 3.2
How many people were in each class on the Titanic? From this display, it looks as though the service must have been great, since most aboard were crew members. Although the length of each ship here corresponds to the correct number, the impression is all wrong. In fact, only about 40% were crew.

Now that we have the frequency table, we’re ready to follow the three rules of data analysis and make a picture of the data. But a bad picture can distort our understanding rather than help it. Here’s a graph of the *Titanic* data. What impression do you get about who was aboard the ship?

It sure looks like most of the people on the *Titanic* were crew members, with a few passengers along for the ride. That doesn’t seem right. What’s wrong? The lengths of the ships *do* match the totals in the table. (You can check the scale at the bottom.) However, experience and psychological tests show that our eyes tend to be more impressed by the *area* than by other aspects of each ship image. So, even though the *length* of each ship matches up with one of the totals, it’s the associated *area* in the image that we notice. Since there were about 3 times as many crew as second-class passengers, the ship depicting the number of crew is about 3 times longer than the ship depicting second-class passengers, but it occupies about 9 times the area. As you can see from the frequency table (Table 3.2), that just isn’t a correct impression.

The best data displays observe a fundamental principle of graphing data called the **area principle**. The area principle says that the area occupied by a part of the graph should correspond to the magnitude of the value it represents. Violations of the area principle are a common way to lie (or, since most mistakes are unintentional, we should say err) with Statistics.

Bar Charts

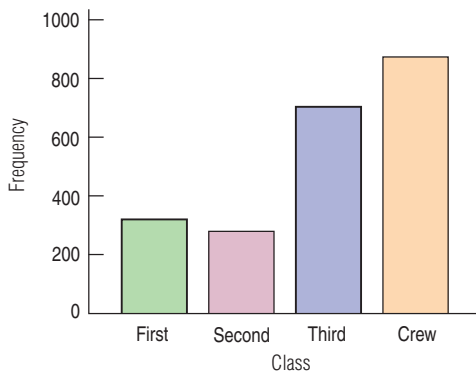
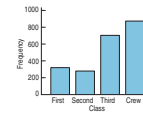


FIGURE 3.3 *People on the Titanic by Ticket Class*
With the area principle satisfied, we can see the true distribution more clearly.

Here’s a chart that obeys the area principle. It’s not as visually entertaining as the ships, but it does give an *accurate* visual impression of the distribution. The height of each bar shows the count for its category. The bars are the same width, so their heights determine their areas, and the areas are proportional to the counts in each class. Now it’s easy to see that the majority of people on board were *not* crew, as the ships picture led us to believe. We can also see that there were about 3 times as many crew as second-class passengers. And there were more than twice as many third-class passengers as either first- or second-class passengers, something you may have missed in the frequency table. Bar charts make these kinds of comparisons easy and natural.

A **bar chart** displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison. Bar charts should have small spaces between the bars to indicate that these are freestanding bars that could be rearranged into any order. The bars are lined up along a common base.

Usually they stick up like this



but sometimes they run

sideways like this



If we really want to draw attention to the relative *proportion* of passengers falling into each of these classes, we could replace the counts with percentages and use a **relative frequency bar chart**.

AS Activity: Bar Charts.

Watch bar charts grow from data; then use your statistics package to create some bar charts for yourself.

For some reason, some computer programs give the name “bar chart” to any graph that uses bars. And others use different names according to whether the bars are horizontal or vertical. Don’t be misled. “Bar chart” is the term for a *display of counts of a categorical variable with bars*.

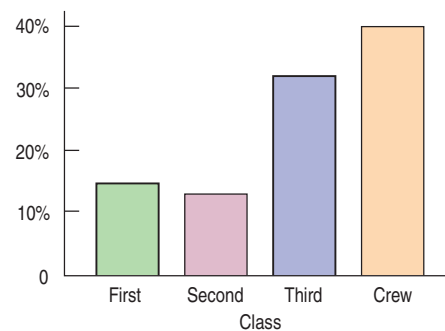


FIGURE 3.4

The relative frequency bar chart looks the same as the bar chart (Figure 3.3) but shows the proportion of people in each category rather than the counts.

Pie Charts

Another common display that shows how a whole group breaks into several categories is a pie chart. **Pie charts** show the whole group of cases as a circle. They slice the circle into pieces whose sizes are proportional to the fraction of the whole in each category.

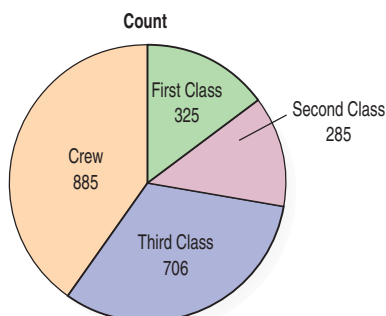


FIGURE 3.5 Number of Titanic passengers in each class

Pie charts give a quick impression of how a whole group is partitioned into smaller groups. Because we’re used to cutting up pies into 2, 4, or 8 pieces, pie charts are good for seeing relative frequencies near $1/2$, $1/4$, or $1/8$. For example, you may be able to tell that the pink slice, representing the second-class passengers, is very close to $1/8$ of the total. It’s harder to see that there were about twice as many third-class as first-class passengers. Which category had the most passengers? Were there more crew or more third-class passengers? Comparisons such as these are easier in a bar chart.

Think before you draw. Our first rule of data analysis is *Make a picture*. But what kind of picture? We don’t have a lot of options—yet. There’s more to Statistics than pie charts and bar charts, and knowing when to use each type of graph is a critical first step in data analysis. That decision depends in part on what type of data we have.

It’s important to check that the data are appropriate for whatever method of analysis you choose. Before you make a bar chart or a pie chart, always check the

Categorical Data Condition: The data are counts or percentages of individuals in categories.

If you want to make a relative frequency bar chart or a pie chart, you'll need to also make sure that the categories don't overlap so that no individual is counted twice. If the categories do overlap, you can still make a bar chart, but the percentages won't add up to 100%. For the *Titanic* data, either kind of display is appropriate because the categories don't overlap.

Throughout this course, you'll see that doing Statistics right means selecting the proper methods. That means you have to *Think* about the situation at hand. An important first step, then, is to check that the type of analysis you plan is appropriate. The Categorical Data Condition is just the first of many such checks.

Contingency Tables: Children and First-Class Ticket Holders First?

AS **Activity: Children at Risk.** This activity looks at the fates of children aboard the *Titanic*; the subsequent activity shows how to make such tables on a computer.

We know how many tickets of each class were sold on the *Titanic*, and we know that only about 32% of all those aboard the *Titanic* survived. After looking at the distribution of each variable by itself, it's natural and more interesting to ask how they relate. Was there a relationship between the kind of ticket a passenger held and the passenger's chances of making it into the lifeboat? To answer this question, we need to look at the two categorical variables *Class* and *Survival* together.

To look at two categorical variables together, we often arrange the counts in a two-way table. Here is a two-way table of those aboard the *Titanic*, classified according to the class of ticket and whether the ticket holder survived or didn't. Because the table shows how the individuals are distributed along each variable, contingent on the value of the other variable, such a table is called a **contingency table**.

Contingency table of ticket *Class* and *Survival*. The bottom line of "Totals" is the same as the previous frequency table.

Table 3.4

		Class				Total
		First	Second	Third	Crew	
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

The margins of the table, both on the right and at the bottom, give totals. The bottom line of the table is just the frequency distribution of ticket *Class*. The right column of the table is the frequency distribution of the variable *Survival*. When presented like this, in the margins of a contingency table, the frequency distribution of one of the variables is called its **marginal distribution**.

Each **cell** of the table gives the count for a combination of values of the two variables. If you look down the column for second-class passengers to the first cell, you can see that 118 second-class passengers survived. Looking at the third-class passengers, you can see that more third-class passengers (178) survived. Were second-class passengers more likely to survive? Questions like this are easier to address by using percentages. The 118 survivors in second class were 41.4% of the total 285 second-class passengers, while the 178 surviving third-class passengers were only 25.2% of that class's total.

We know that 118 second-class passengers survived. We could display this number as a percentage—but as a percentage of what? The total number of passengers? (118 is 5.4% of the total: 2201.) The number of second-class passengers?



A bell-shaped artifact from the *Titanic*.

(118 is 41.4% of the 285 second-class passengers.) The number of survivors? (118 is 16.6% of the 711 survivors.) All of these are possibilities, and all are potentially useful or interesting. You'll probably wind up calculating (or letting your technology calculate) lots of percentages. Most statistics programs offer a choice of total percent, row percent, or column percent for contingency tables. Unfortunately, they often put them all together with several numbers in each cell of the table. The resulting table holds lots of information, but it can be hard to understand:

Another contingency table of ticket Class. This time we see not only the counts for each combination of *Class* and *Survival* (in bold) but the percentages these counts represent. For each count, there are three choices for the percentage: by row, by column, and by table total. There's probably too much information here for this table to be useful.

Table 3.5

		Class					Total
		First	Second	Third	Crew		
Survival	Alive	Count	203	118	178	212	711
		% of Row	28.6%	16.6%	25.0%	29.8%	100%
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
		% of Table	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Row	8.2%	11.2%	35.4%	45.2%	100%
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
		% of Table	5.6%	7.6%	24.0%	30.6%	67.7%
	Total	Count	325	285	706	885	2201
		% of Row	14.8%	12.9%	32.1%	40.2%	100%
		% of Column	100%	100%	100%	100%	100%
		% of Table	14.8%	12.9%	32.1%	40.2%	100%

To simplify the table, let's first pull out the percent of table values:

A contingency table of Class by Survival with only the table percentages

Table 3.6

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	5.6%	7.6%	24.0%	30.6%	67.7%
	Total	14.8%	12.9%	32.1%	40.2%	100%

These percentages tell us what percent of *all* passengers belong to each combination of column and row category. For example, we see that although 8.1% of the people aboard the *Titanic* were surviving third-class ticket holders, only 5.4% were surviving second-class ticket holders. Is this fact useful? Comparing these percentages, you might think that the chances of surviving were better in third class than in second. But be careful. There were many more third-class than second-class passengers on the *Titanic*, so there were more third-class survivors. That group is a larger percentage of the passengers, but is that really what we want to know?

Percent of what? The English language can be tricky when we talk about percentages. If you're asked "What percent of the survivors were in second class?" it's pretty clear that we're interested only in survivors. It's as if we're restricting the *Who* in the question to the survivors, so we should look at the number of second-class passengers among all the survivors—in other words, the row percent. But if you're asked "What percent were second-class passengers who survived?" you have a different question. Be careful; here, the *Who* is everyone on board, so 2201 should be the denominator, and the answer is the table percent.

And if you're asked "What percent of the second-class passengers survived?" you have a third question. Now the *Who* is the second-class passengers, so the denominator is the 285 second-class passengers, and the answer is the column percent.

Always be sure to ask "percent of what?" That will help you to know the *Who* and whether we want *row*, *column*, or *table* percentages.

FOR EXAMPLE

Finding marginal distributions

In January 2007, a Gallup poll asked 1008 Americans age 18 and over whether they planned to watch the upcoming Super Bowl. The pollster also asked those who planned to watch whether they were looking forward more to seeing the football game or the commercials. The results are summarized in the table:

Question: What's the marginal distribution of the responses?

To determine the percentages for the three responses, divide the count for each response by the total number of people polled:

$$\frac{479}{1008} = 47.5\% \quad \frac{237}{1008} = 23.5\% \quad \frac{292}{1008} = 29.0\%$$

According to the poll, 47.5% of American adults were looking forward to watching the Super Bowl game, 23.5% were looking forward to watching the commercials, and 29% didn't plan to watch at all.

		Sex		
		Male	Female	Total
Response	Game	279	200	479
	Commercials	81	156	237
	Won't watch	132	160	292
Total		492	516	1008

Conditional Distributions

The more interesting questions are *contingent*. We'd like to know, for example, what percentage of *second-class passengers* survived and how that compares with the survival rate for third-class passengers.

It's more interesting to ask whether the chance of surviving the *Titanic* sinking *depended* on ticket class. We can look at this question in two ways. First, we could ask how the distribution of ticket *Class* changes between survivors and non-survivors. To do that, we look at the *row percentages*:

The conditional distribution of ticket *Class* conditioned on each value of *Survival*: *Alive* and *Dead*.

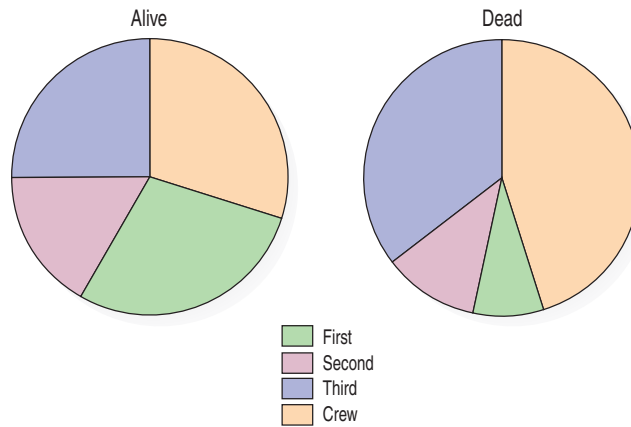
Table 3.7

		Class				Total
		First	Second	Third	Crew	
Survival	Alive	203 28.6%	118 16.6%	178 25.0%	212 29.8%	711 100%
	Dead	122 8.2%	167 11.2%	528 35.4%	673 45.2%	1490 100%

By focusing on each row separately, we see the distribution of class under the *condition* of surviving or not. The sum of the percentages in each row is 100%, and we divide that up by ticket class. In effect, we temporarily restrict the *Who* first to survivors and make a pie chart for them. Then we refocus the *Who* on the nonsurvivors and make their pie chart. These pie charts show the distribution of ticket classes *for each row* of the table: survivors and nonsurvivors. The distributions we create this way are called **conditional distributions**, because they show the distribution of one variable for just those cases that satisfy a condition on another variable.

FIGURE 3.6

Pie charts of the conditional distributions of ticket Class for the survivors and nonsurvivors, separately. Do the distributions appear to be the same? We're primarily concerned with percentages here, so pie charts are a reasonable choice.



FOR EXAMPLE Finding conditional distributions

Recap: The table shows results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn't plan to watch.

Question: How do the conditional distributions of interest in the commercials differ for men and women?

		Sex		Total
		Male	Female	
Response	Game	279	200	479
	Commercials	81	156	237
	Won't watch	132	160	292
	Total	492	516	1008

Look at the group of people who responded "Commercials" and determine what percent of them were male and female:

$$\frac{81}{237} = 34.2\% \quad \frac{156}{237} = 65.8\%$$

Women make up a sizable majority of the adult Americans who look forward to seeing Super Bowl commercials more than the game itself. Nearly 66% of people who voiced a preference for the commercials were women, and only 34% were men.

But we can also turn the question around. We can look at the distribution of *Survival* for each category of ticket *Class*. To do this, we look at the *column percentages*. Those show us whether the chance of surviving was roughly the same for each of the four classes. Now the percentages in each column add to 100%, because we've restricted the *Who*, in turn, to each of the four ticket classes:

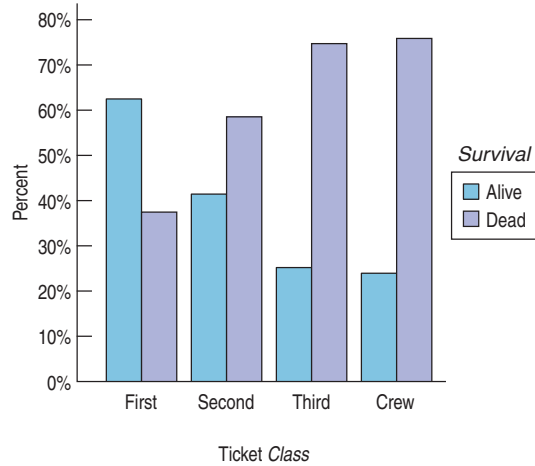
A contingency table of *Class* by *Survival* with only counts and column percentages. Each column represents the conditional distribution of *Survival* for a given category of ticket *Class*.

Table 3.8

		Class				Total	
		First	Second	Third	Crew		
Survival	Alive	Count % of Column	203 62.5%	118 41.4%	178 25.2%	212 24.0%	711 32.3%
	Dead	Count % of Column	122 37.5%	167 58.6%	528 74.8%	673 76.0%	1490 67.7%
	Total	Count	325 100%	285 100%	706 100%	885 100%	2201 100%

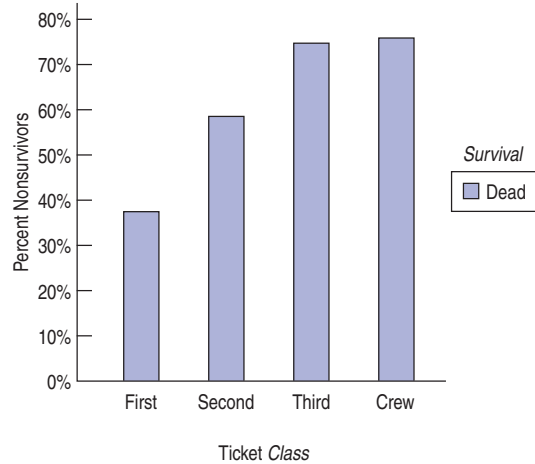
Looking at how the percentages change across each row, it sure looks like ticket class mattered in whether a passenger survived. To make it more vivid, we could show the distribution of *Survival* for each ticket class in a display. Here's a side-by-side bar chart showing percentages of surviving and not for each category:

FIGURE 3.7
Side-by-side bar chart showing the conditional distribution of *Survival* for each category of ticket *Class*. The corresponding pie charts would have only two categories in each of four pies, so bar charts seem the better alternative.



These bar charts are simple because, for the variable *Survival*, we have only two alternatives: Alive and Dead. When we have only two categories, we really need to know only the percentage of one of them. Knowing the percentage that survived tells us the percentage that died. We can use this fact to simplify the display even more by dropping one category. Here are the percentages of dying across the classes displayed in one chart:

FIGURE 3.8
 Bar chart showing just nonsurvivor percentages for each value of ticket *Class*. Because we have only two values, the second bar doesn't add any information. Compare this chart to the side-by-side bar chart shown earlier.



TI-*n*spire
Conditional distributions and association. Explore the *Titanic* data to see which passengers were most likely to survive.

Now it's easy to compare the risks. Among first-class passengers, 37.5% perished, compared to 58.6% for second-class ticket holders, 74.8% for those in third class, and 76.0% for crew members.

If the risk had been about the same across the ticket classes, we would have said that survival was *independent* of class. But it's not. The differences we see among these conditional distributions suggest that survival may have depended on ticket class. You may find it useful to consider conditioning on each variable in a contingency table in order to explore the dependence between them.

It is interesting to know that *Class* and *Survival* are associated. That’s an important part of the *Titanic* story. And we know how important this is because the margins show us the actual numbers of people involved.

Variables can be associated in many ways and to different degrees. The best way to tell whether two variables are associated is to ask whether they are *not*.¹ In a contingency table, when the distribution of *one* variable is the same for all categories of another, we say that the variables are **independent**. That tells us there’s no association between these variables. We’ll see a way to check for independence formally later in the book. For now, we’ll just compare the distributions.

FOR EXAMPLE

Looking for associations between variables

Recap: The table shows results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn’t plan to watch.

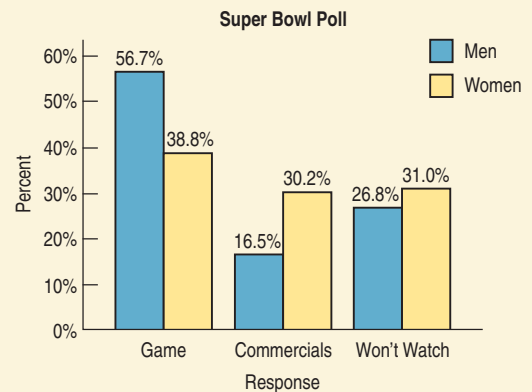
Question: Does it seem that there’s an association between interest in Super Bowl TV coverage and a person’s sex?

		Sex		
		Male	Female	Total
Response	Game	279	200	479
	Commercials	81	156	237
	Won’t watch	132	160	292
	Total	492	516	1008

First find the distribution of the three responses for the men (the column percentages):

$$\frac{279}{492} = 56.7\% \quad \frac{81}{492} = 16.5\% \quad \frac{132}{492} = 26.8\%$$

Then do the same for the women who were polled, and display the two distributions with a side-by-side bar chart:



Based on this poll it appears that women were only slightly less interested than men in watching the Super Bowl tele-cast: 31% of the women said they didn’t plan to watch, compared to just under 27% of men. Among those who planned to watch, however, there appears to be an association between the viewer’s sex and what the viewer is most looking forward to. While more women are interested in the game (39%) than the commercials (30%), the margin among men is much wider: 57% of men said they were looking forward to seeing the game, compared to only 16.5% who cited the commercials.

¹This kind of “backwards” reasoning shows up surprisingly often in science—and in Statistics. We’ll see it again.



JUST CHECKING

A Statistics class reports the following data on Sex and Eye Color for students in the class:

		Eye Color			Total
		Blue	Brown	Green/Hazel/Other	
Sex	Males	6	20	6	32
	Females	4	16	12	32
	Total	10	36	18	64

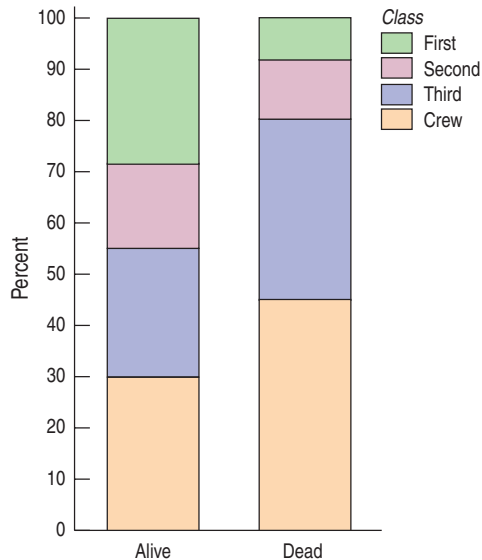
1. What percent of females are brown-eyed?
2. What percent of brown-eyed students are female?
3. What percent of students are brown-eyed females?
4. What's the distribution of Eye Color?
5. What's the conditional distribution of Eye Color for the males?
6. Compare the percent who are female among the blue-eyed students to the percent of all students who are female.
7. Does it seem that Eye Color and Sex are independent? Explain.

Segmented Bar Charts

We could display the *Titanic* information by dividing up bars rather than circles. The resulting **segmented bar chart** treats each bar as the “whole” and divides it proportionally into segments corresponding to the percentage in each group. We can clearly see that the distributions of ticket *Class* are different, indicating again that survival was not independent of ticket *Class*.

FIGURE 3.9 A segmented bar chart for Class by Survival

Notice that although the totals for survivors and nonsurvivors are quite different, the bars are the same height because we have converted the numbers to percentages. Compare this display with the side-by-side pie charts of the same data in Figure 3.6.



STEP-BY-STEP EXAMPLE

Examining Contingency Tables

Medical researchers followed 6272 Swedish men for 30 years to see if there was any association between the amount of fish in their diet and prostate cancer (“Fatty Fish Consumption and Risk of Prostate Cancer,” *Lancet*, June 2001). Their results are summarized in this table:



We asked for a picture of a man eating fish. This is what we got.

		Prostate Cancer	
		No	Yes
Fish Consumption	Never/seldom	110	14
	Small part of diet	2420	201
	Moderate part	2769	209
	Large part	507	42

Table 3.9

Question: Is there an association between fish consumption and prostate cancer?



Plan Be sure to state what the problem is about.

Variables Identify the variables and report the W’s.

Be sure to check the appropriate condition.

I want to know if there is an association between fish consumption and prostate cancer.

The individuals are 6272 Swedish men followed by medical researchers for 30 years. The variables record their fish consumption and whether or not they were diagnosed with prostate cancer.

✓ **Categorical Data Condition:** I have counts for both fish consumption and cancer diagnosis. The categories of diet do not overlap, and the diagnoses do not overlap. It’s okay to draw pie charts or bar charts.

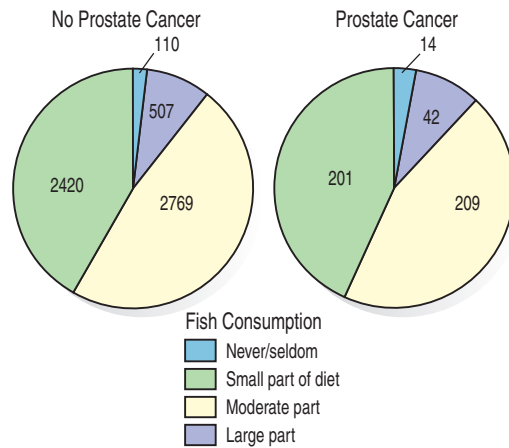


Mechanics It’s a good idea to check the marginal distributions first before looking at the two variables together.

		Prostate Cancer		
		No	Yes	Total
Fish Consumption	Never/seldom	110	14	124 (2.0%)
	Small part of diet	2420	201	2621 (41.8%)
	Moderate part	2769	209	2978 (47.5%)
	Large part	507	42	549 (8.8%)
	Total	5806 (92.6%)	466 (7.4%)	6272 (100%)

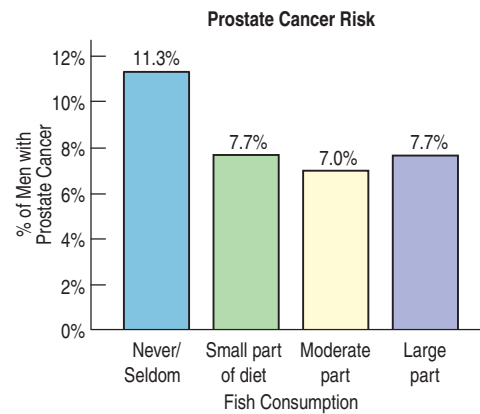
Two categories of the diet are quite small, with only 2.0% Never/Seldom eating fish and 8.8% in the “Large part” category. Overall, 7.4% of the men in this study had prostate cancer.

Then, make appropriate displays to see whether there is a difference in the relative proportions. These pie charts compare fish consumption for men who have prostate cancer to fish consumption for men who don't.



It's hard to see much difference in the pie charts. So, I made a display of the row percentages. Because there are only two alternatives, I chose to display the risk of prostate cancer for each group:

Both pie charts and bar charts can be used to compare conditional distributions. Here we compare prostate cancer rates based on differences in fish consumption.



Conclusion Interpret the patterns in the table and displays in context. If you can, discuss possible real-world consequences. Be careful not to overstate what you see. The results may not generalize to other situations.

Overall, there is a 7.4% rate of prostate cancer among men in this study. Most of the men (89.3%) ate fish either as a moderate or small part of their diet. From the pie charts, it's hard to see a difference in cancer rates among the groups. But in the bar chart, it looks like the cancer rate for those who never/seldom ate fish may be somewhat higher.

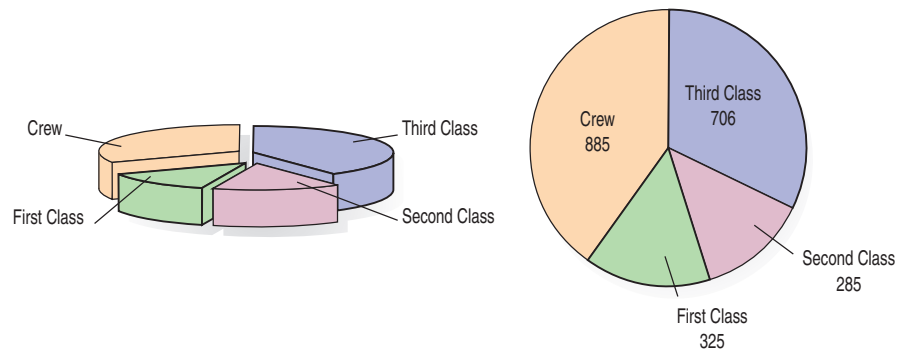
However, only 124 of the 6272 men in the study fell into this category, and only 14 of them developed prostate cancer. More study would probably be needed before we would recommend that men change their diets.²

²The original study actually used pairs of twins, which enabled the researchers to discern that the risk of cancer for those who never ate fish actually *was* substantially greater. Using pairs is a special way of gathering data. We'll discuss such study design issues and how to analyze the data in the later chapters.

This study is an example of looking at a sample of data to learn something about a larger population. We care about more than these particular 6272 Swedish men. We hope that learning about their experiences will tell us something about the value of eating fish in general. That raises the interesting question of what population we think this sample might represent. Do we hope to learn about all Swedish men? About all men? About the value of eating fish for all adult humans? ³ Often, it can be hard to decide just which population our findings may tell us about, but that also is how researchers decide what to look into in future studies.

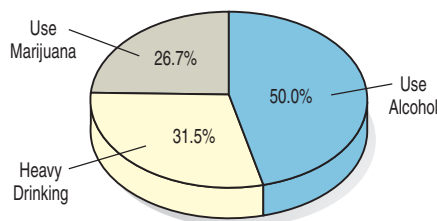
WHAT CAN GO WRONG?

- ▶ **Don't violate the area principle.** This is probably the most common mistake in a graphical display. It is often made in the cause of artistic presentation. Here, for example, are two displays of the pie chart of the *Titanic* passengers by class:



The one on the left looks pretty, doesn't it? But showing the pie on a slant violates the area principle and makes it much more difficult to compare fractions of the whole made up of each class—the principal feature that a pie chart ought to show.

- ▶ **Keep it honest.** Here's a pie chart that displays data on the percentage of high school students who engage in specified dangerous behaviors as reported by the Centers for Disease Control. What's wrong with this plot?

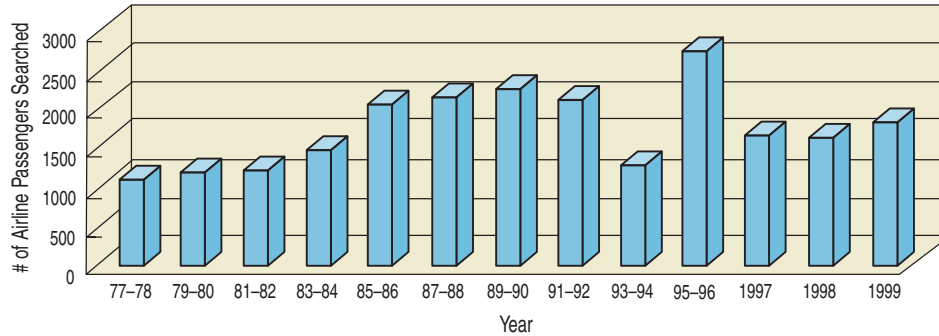


Try adding up the percentages. Or look at the 50% slice. Does it look right? Then think: What are these percentages of? Is there a "whole" that has been sliced up? In a pie chart, the proportions shown by each slice of the pie must add up to 100% and each individual must fall into only one category. Of course, showing the pie on a slant makes it even harder to detect the error.

(continued)

³ Probably not, since we're looking only at prostate cancer risk.

Here’s another. This bar chart shows the number of airline passengers searched in security screening, by year:



Looks like things didn’t change much in the final years of the 20th century—until you read the bar labels and see that the last three bars represent single years while all the others are for *pairs* of years. Of course, the false depth makes it harder to see the problem.

- ▶ **Don’t confuse similar-sounding percentages.** These percentages sound similar but are different:
 - ▶ The percentage of the passengers who were both in first class and survived: This would be $203/2201$, or 9.4%.
 - ▶ The percentage of the first-class passengers who survived: This is $203/325$, or 62.5%.
 - ▶ The percentage of the survivors who were in first class: This is $203/711$, or 28.6%.

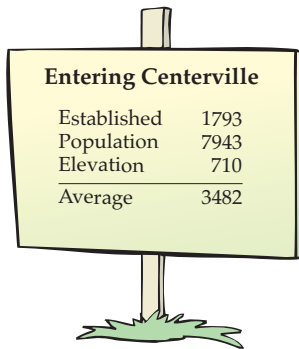
In each instance, pay attention to the *Who* implicitly defined by the phrase. Often there is a restriction to a smaller group (all aboard the *Titanic*, those in first class, and those who survived, respectively) before a percentage is found. Your discussion of results must make these differences clear.

- ▶ **Don’t forget to look at the variables separately, too.** When you make a contingency table or display a conditional distribution, be sure you also examine the marginal distributions. It’s important to know how many cases are in each category.
- ▶ **Be sure to use enough individuals.** When you consider percentages, take care that they are based on a large enough number of individuals. Take care not to make a report such as this one:

We found that 66.67% of the rats improved their performance with training. The other rat died.

- ▶ **Don’t overstate your case.** Independence is an important concept, but it is rare for two variables to be *entirely* independent. We can’t conclude that one variable has no effect whatsoever on another. Usually, all we know is that little effect was observed in our study. Other studies of other groups under other circumstances could find different results.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201



SIMPSON’S PARADOX

- ▶ **Don’t use unfair or silly averages.** Sometimes averages can be misleading. Sometimes they just don’t make sense at all. Be careful when averaging different variables that the quantities you’re averaging are comparable. The Centerville sign says it all.

When using averages of proportions across several different groups, it’s important to make sure that the groups really are comparable.

It's easy to make up an example showing that averaging across very different values or groups can give absurd results. Here's how that might work: Suppose there are two pilots, Moe and Jill. Moe argues that he's the better pilot of the two, since he managed to land 83% of his last 120 flights on time compared with Jill's 78%. But let's look at the data a little more closely. Here are the results for each of their last 120 flights, broken down by the time of day they flew:

Table 3.10

On-time flights by *Time of Day* and *Pilot*. Look at the percentages within each *Time of Day* category. Who has a better on-time record during the day? At night? Who is better overall?

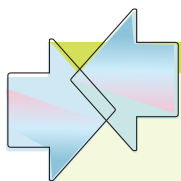
		Time of Day		
		Day	Night	Overall
Pilot	Moe	90 out of 100 90%	10 out of 20 50%	100 out of 120 83%
	Jill	19 out of 20 95%	75 out of 100 75%	94 out of 120 78%

One famous example of Simpson's paradox arose during an investigation of admission rates for men and women at the University of California at Berkeley's graduate schools. As reported in an article in *Science*, about 45% of male applicants were admitted, but only about 30% of female applicants got in. It looked like a clear case of discrimination. However, when the data were broken down by school (Engineering, Law, Medicine, etc.), it turned out that, within each school, the women were admitted at nearly the same or, in some cases, much *higher* rates than the men. How could this be? Women applied in large numbers to schools with very low admission rates (Law and Medicine, for example, admitted fewer than 10%). Men tended to apply to Engineering and Science. Those schools have admission rates above 50%. When the *average* was taken, the women had a much lower *overall* rate, but the average didn't really make sense.

Look at the daytime and nighttime flights separately. For day flights, Jill had a 95% on-time rate and Moe only a 90% rate. At night, Jill was on time 75% of the time and Moe only 50%. So Moe is better "overall," but Jill is better both during the day and at night. How can this be?

What's going on here is a problem known as **Simpson's paradox**, named for the statistician who discovered it in the 1960s. It comes up rarely in real life, but there have been several well-publicized cases. As we can see from the pilot example, the problem is *unfair averaging* over different groups. Jill has mostly night flights, which are more difficult, so her *overall average* is heavily influenced by her nighttime average. Moe, on the other hand, benefits from flying mostly during the day, with its higher on-time percentage. With their very different patterns of flying conditions, taking an overall average is misleading. It's not a fair comparison.

The moral of Simpson's paradox is to be careful when you average across different levels of a second variable. It's always better to compare percentages or other averages *within* each level of the other variable. The overall average may be misleading.



CONNECTIONS

All of the methods of this chapter work with *categorical variables*. You must know the *Who* of the data to know who is counted in each category and the *What* of the variable to know where the categories come from.



WHAT HAVE WE LEARNED?

We've learned that we can summarize categorical data by counting the number of cases in each category, sometimes expressing the resulting distribution as percents. We can display the distribution in a bar chart or a pie chart. When we want to see how two categorical variables are related, we put the counts (and/or percentages) in a two-way table called a contingency table.

- ▶ We look at the marginal distribution of each variable (found in the margins of the table).
- ▶ We also look at the conditional distribution of a variable within each category of the other variable.
- ▶ We can display these conditional and marginal distributions by using bar charts or pie charts.
- ▶ If the conditional distributions of one variable are (roughly) the same for every category of the other, the variables are independent.

Terms

Frequency table
(Relative frequency table)

Distribution

Area principle

Bar chart
(Relative frequency bar chart)

Pie chart

Categorical data condition

Contingency table

Marginal distribution

Conditional distribution

Independence

Segmented bar chart

Simpson's paradox

21. A frequency table lists the categories in a categorical variable and gives the count (or percentage of observations for each category.

22. The distribution of a variable gives

- ▶ the possible values of the variable and
- ▶ the relative frequency of each value.

22. In a statistical display, each data value should be represented by the same amount of area.

22. Bar charts show a bar whose area represents the count (or percentage) of observations for each category of a categorical variable.

23. Pie charts show how a "whole" divides into categories by showing a wedge of a circle whose area corresponds to the proportion in each category.

24. The methods in this chapter are appropriate for displaying and describing categorical data. Be careful not to use them with quantitative data.

24. A contingency table displays counts and, sometimes, percentages of individuals falling into named categories on two or more variables. The table categorizes the individuals on all variables at once to reveal possible patterns in one variable that may be contingent on the category of the other.

24. In a contingency table, the distribution of either variable alone is called the marginal distribution. The counts or percentages are the totals found in the margins (last row or column) of the table.

26. The distribution of a variable restricting the *Who* to consider only a smaller group of individuals is called a conditional distribution.

29. Variables are said to be independent if the conditional distribution of one variable is the same for each category of the other. We'll show how to check for independence in a later chapter.

30. A segmented bar chart displays the conditional distribution of a categorical variable within each category of another variable.

34. When averages are taken across different groups, they can appear to contradict the overall averages. This is known as "Simpson's paradox."

Skills

THINK

- ▶ Be able to recognize when a variable is categorical and choose an appropriate display for it.
- ▶ Understand how to examine the association between categorical variables by comparing conditional and marginal percentages.

SHOW

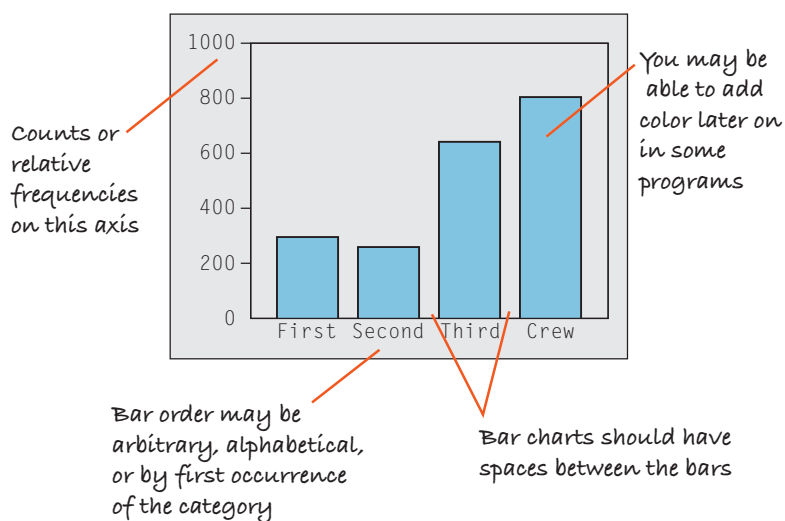
- ▶ Be able to summarize the distribution of a categorical variable with a frequency table.
- ▶ Be able to display the distribution of a categorical variable with a bar chart or pie chart.
- ▶ Know how to make and examine a contingency table.



- ▶ Know how to make and examine displays of the conditional distributions of one variable for two or more groups.
- ▶ Be able to describe the distribution of a categorical variable in terms of its possible values and relative frequencies.
- ▶ Know how to describe any anomalies or extraordinary features revealed by the display of a variable.
- ▶ Be able to describe and discuss patterns found in a contingency table and associated displays of conditional distributions.

DISPLAYING CATEGORICAL DATA ON THE COMPUTER

Although every package makes a slightly different bar chart, they all have similar features:



Sometimes the count or a percentage is printed above or on top of each bar to give some additional information. You may find that your statistics package sorts category names in annoying orders by default. For example, many packages sort categories alphabetically or by the order the categories are seen in the data set. Often, neither of these is the best choice.

EXERCISES

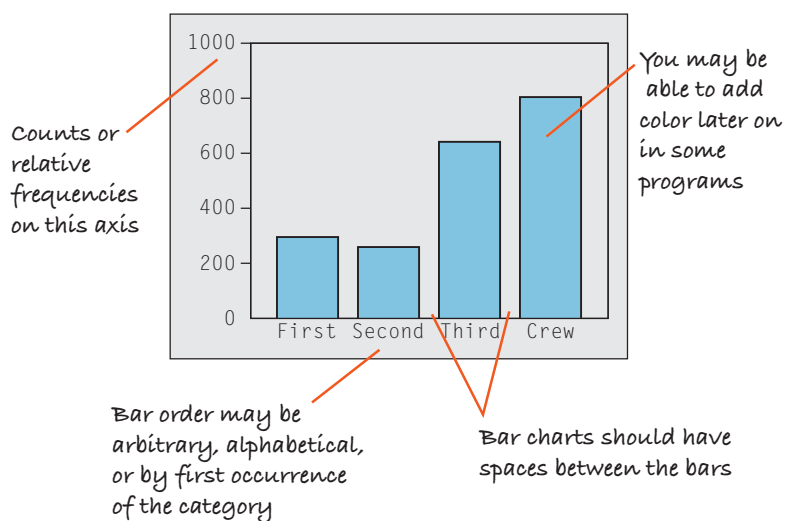
1. **Graphs in the news.** Find a bar graph of categorical data from a newspaper, a magazine, or the Internet.
 - a) Is the graph clearly labeled?
 - b) Does it violate the area principle?
 - c) Does the accompanying article tell the W's of the variable?
 - d) Do you think the article correctly interprets the data? Explain.
2. **Graphs in the news II.** Find a pie chart of categorical data from a newspaper, a magazine, or the Internet.
 - a) Is the graph clearly labeled?
 - b) Does it violate the area principle?
 - c) Does the accompanying article tell the W's of the variable?
 - d) Do you think the article correctly interprets the data? Explain.



- ▶ Know how to make and examine displays of the conditional distributions of one variable for two or more groups.
- ▶ Be able to describe the distribution of a categorical variable in terms of its possible values and relative frequencies.
- ▶ Know how to describe any anomalies or extraordinary features revealed by the display of a variable.
- ▶ Be able to describe and discuss patterns found in a contingency table and associated displays of conditional distributions.

DISPLAYING CATEGORICAL DATA ON THE COMPUTER

Although every package makes a slightly different bar chart, they all have similar features:



Sometimes the count or a percentage is printed above or on top of each bar to give some additional information. You may find that your statistics package sorts category names in annoying orders by default. For example, many packages sort categories alphabetically or by the order the categories are seen in the data set. Often, neither of these is the best choice.

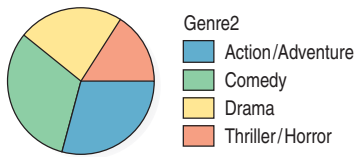
EXERCISES

1. **Graphs in the news.** Find a bar graph of categorical data from a newspaper, a magazine, or the Internet.
 - a) Is the graph clearly labeled?
 - b) Does it violate the area principle?
 - c) Does the accompanying article tell the W's of the variable?
 - d) Do you think the article correctly interprets the data? Explain.
2. **Graphs in the news II.** Find a pie chart of categorical data from a newspaper, a magazine, or the Internet.
 - a) Is the graph clearly labeled?
 - b) Does it violate the area principle?
 - c) Does the accompanying article tell the W's of the variable?
 - d) Do you think the article correctly interprets the data? Explain.

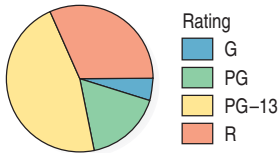
3. **Tables in the news.** Find a frequency table of categorical data from a newspaper, a magazine, or the Internet.
- Is it clearly labeled?
 - Does it display percentages or counts?
 - Does the accompanying article tell the *W*'s of the variable?
 - Do you think the article correctly interprets the data? Explain.

4. **Tables in the news II.** Find a contingency table of categorical data from a newspaper, a magazine, or the Internet.
- Is it clearly labeled?
 - Does it display percentages or counts?
 - Does the accompanying article tell the *W*'s of the variables?
 - Do you think the article correctly interprets the data? Explain.

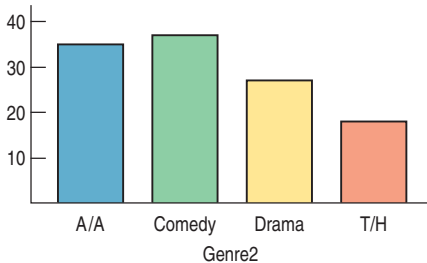
- T** 5. **Movie genres.** The pie chart summarizes the genres of 120 first-run movies released in 2005.
- Is this an appropriate display for the genres? Why/why not?
 - Which genre was least common?



- T** 6. **Movie ratings.** The pie chart shows the ratings assigned to 120 first-run movies released in 2005.
- Is this an appropriate display for these data? Explain.
 - Which was the most common rating?

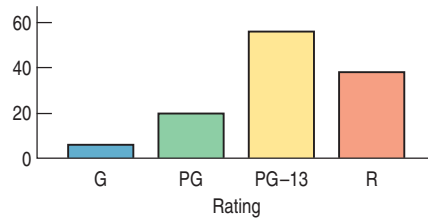


- T** 7. **Genres again.** Here is a bar chart summarizing the 2005 movie genres, as seen in the pie chart in Exercise 5.
- Which genre was most common?
 - Is it easier to see that in the pie chart or the bar chart? Explain.



- T** 8. **Ratings again.** Here is a bar chart summarizing the 2005 movie ratings, as seen in the pie chart in Exercise 6.
- Which was the least common rating?
 - An editorial claimed that there's been a growth in PG-13 rated films that, according to the writer, "have too much sex and violence," at the expense of G-rated

films that offer "good, clean fun." The writer offered the bar chart below as evidence to support his claim. Does the bar chart support his claim? Explain.



9. **Magnet schools.** An article in the Winter 2003 issue of *Chance* magazine reported on the Houston Independent School District's magnet schools programs. Of the 1755 qualified applicants, 931 were accepted, 298 were wait-listed, and 526 were turned away for lack of space. Find the relative frequency distribution of the decisions made, and write a sentence describing it.
10. **Magnet schools again.** The *Chance* article about the Houston magnet schools program described in Exercise 9 also indicated that 517 applicants were black or Hispanic, 292 Asian, and 946 white. Summarize the relative frequency distribution of ethnicity with a sentence or two (in the proper context, of course).
11. **Causes of death 2004.** The Centers for Disease Control and Prevention (www.cdc.gov) lists causes of death in the United States during 2004:

Cause of Death	Percent
Heart disease	27.2
Cancer	23.1
Circulatory diseases and stroke	6.3
Respiratory diseases	5.1
Accidents	4.7

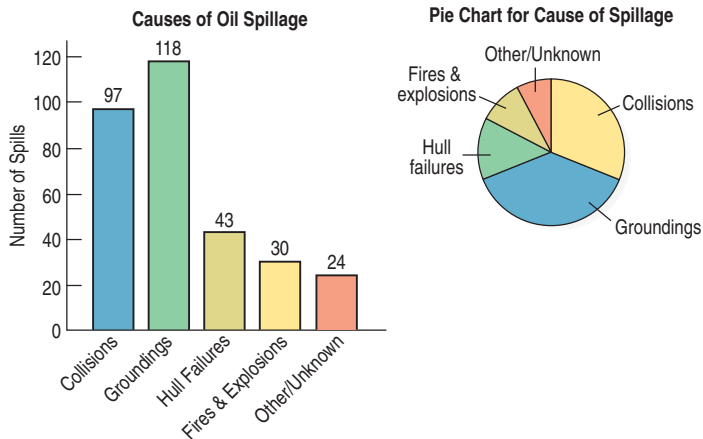
- Is it reasonable to conclude that heart or respiratory diseases were the cause of approximately 33% of U.S. deaths in 2003?
 - What percent of deaths were from causes not listed here?
 - Create an appropriate display for these data.
12. **Plane crashes.** An investigation compiled information about recent nonmilitary plane crashes (www.planecrashinfo.com). The causes, to the extent that they could be determined, are summarized in the table.

Cause	Percent
Pilot error	40
Other human error	5
Weather	6
Mechanical failure	14
Sabotage	6

- Is it reasonable to conclude that the weather or mechanical failures caused only about 20% of recent plane crashes?
- In what percent of crashes were the causes not determined?
- Create an appropriate display for these data.

13. **Oil spills 2006.** Data from the International Tanker Owners Pollution Federation Limited (www.itopf.com) give the cause of spillage for 312 large oil tanker accidents from 1974–2006. Here are displays.

- Write a brief report interpreting what the displays show.
- Is a pie chart an appropriate display for these data? Why or why not?

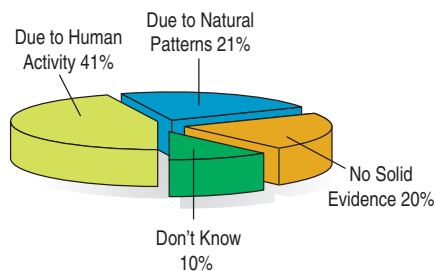


14. **Winter Olympics 2006.** Twenty-six countries won medals in the 2006 Winter Olympics. The table lists them, along with the total number of medals each won:

Country	Medals	Country	Medals
Germany	29	Finland	9
United States	25	Czech Republic	4
Canada	24	Estonia	3
Austria	23	Croatia	3
Russia	22	Australia	2
Norway	19	Poland	2
Sweden	14	Ukraine	2
Switzerland	14	Japan	1
South Korea	11	Belarus	1
Italy	11	Bulgaria	1
China	11	Great Britain	1
France	9	Slovakia	1
Netherlands	9	Latvia	1

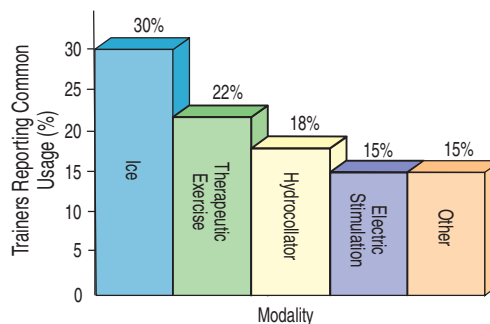
- Try to make a display of these data. What problems do you encounter?
- Can you find a way to organize the data so that the graph is more successful?

15. **Global Warming.** The Pew Research Center for the People and the Press (<http://people-press.org>) has asked a representative sample of U.S. adults about global warming, repeating the question over time. In January 2007, the responses reflected an increased belief that global warming is real and due to human activity. Here's a display of the percentages of respondents choosing each of the major alternatives offered:



List the errors in this display.

16. **Modalities.** A survey of athletic trainers (Scott F. Nadler, Michael Prybicien, Gerard A. Malanga, and Dan Sicher. "Complications from Therapeutic Modalities: Results of a National Survey of Athletic Trainers." *Archives of Physical Medical Rehabilitation* 84 [June 2003]) asked what modalities (treatment methods such as ice, whirlpool, ultrasound, or exercise) they commonly use to treat injuries. Respondents were each asked to list three modalities. The article included the following figure reporting the modalities used:



- What problems do you see with the graph?
- Consider the percentages for the named modalities. Do you see anything odd about them?

17. **Teen smokers.** The organization Monitoring the Future (www.monitoringthefuture.org) asked 2048 eighth graders who said they smoked cigarettes what brands they preferred. The table below shows brand preferences for two regions of the country. Write a few sentences describing the similarities and differences in brand preferences among eighth graders in the two regions listed.

Brand preference	South	West
Marlboro	58.4%	58.0%
Newport	22.5%	10.1%
Camel	3.3%	9.5%
Other (over 20 brands)	9.1%	9.5%
No usual brand	6.7%	12.9%

18. **Handguns.** In an effort to reduce the number of gun-related homicides, some cities have run buyback programs in which the police offer cash (often \$50) to anyone who turns in an operating handgun. *Chance* magazine looked at results from a four-year period in Milwaukee. The table on the next page shows what types of guns were turned in and what types were used in homicides during a four-year period. Write a few sentences comparing the two distributions.

Caliber of gun	Buyback	Homicide
Small (.22, .25, .32)	76.4%	20.3%
Medium (.357, .38, 9 mm)	19.3%	54.7%
Large (.40, .44, .45)	2.1%	10.8%
Other	2.2%	14.2%

T 19. Movies by Genre and Rating. Here's a table that classifies movies released in 2005 by genre and MPAA rating:

	G	PG	PG-13	R	Total
Action/Adventure	66.7	25	30.4	23.7	29.2
Comedy	33.3	60.0	35.7	10.5	31.7
Drama	0	15.0	14.3	44.7	23.3
Thriller/Horror	0	0	19.6	21.1	15.8
Total	100%	100%	100%	100%	100%

- The table gives column percents. How could you tell that from the table itself?
- What percentage of these movies were comedies?
- What percentage of the PG-rated movies were comedies?
- Which of the following can you learn from this table? Give the answer if you can find it from the table.
 - The percentage of PG-13 movies that were comedies
 - The percentage of dramas that were R-rated
 - The percentage of dramas that were G-rated
 - The percentage of 2005 movies that were PG-rated comedies

T 20. The Last Picture Show. Here's another table showing information about 120 movies released in 2005. This table gives percentages of the table total:

	G	PG	PG-13	R	Total
Action/Adventure	3.33%	4.17	14.2	7.50	29.2
Comedy	1.67	10	16.7	3.33	31.7
Drama	0	2.50	6.67	14.2	23.3
Thriller/Horror	0	0	9.17	6.67	15.8
Total	5	16.7	46.7	31.7	100%

- How can you tell that this table holds table percentages (rather than row or column percentages)?
 - What was the most common genre/rating combination in 2005 movies?
 - How many of these movies were PG-rated comedies?
 - How many were G-rated?
 - An editorial about the movies noted, "More than three-quarters of the movies made today can be seen only by patrons 13 years old or older." Does this table support that assertion? Explain.
- 21. Seniors.** Prior to graduation, a high school class was surveyed about its plans. The following table displays the results for white and minority students (the "Minority"

group included African-American, Asian, Hispanic, and Native American students):

Seniors		
	White	Minority
4-year college	198	44
2-year college	36	6
Military	4	1
Employment	14	3
Other	16	3

- What percent of the seniors are white?
 - What percent of the seniors are planning to attend a 2-year college?
 - What percent of the seniors are white and planning to attend a 2-year college?
 - What percent of the white seniors are planning to attend a 2-year college?
 - What percent of the seniors planning to attend a 2-year college are white?
- 22. Politics.** Students in an Intro Stats course were asked to describe their politics as "Liberal," "Moderate," or "Conservative." Here are the results:

Politics					
		L	M	C	Total
Sex	Female	35	36	6	77
	Male	50	44	21	115
	Total	85	80	27	192

- What percent of the class is male?
 - What percent of the class considers themselves to be "Conservative"?
 - What percent of the males in the class consider themselves to be "Conservative"?
 - What percent of all students in the class are males who consider themselves to be "Conservative"?
- 23. More about seniors.** Look again at the table of post-graduation plans for the senior class in Exercise 21.
- Find the conditional distributions (percentages) of plans for the white students.
 - Find the conditional distributions (percentages) of plans for the minority students.
 - Create a graph comparing the plans of white and minority students.
 - Do you see any important differences in the post-graduation plans of white and minority students? Write a brief summary of what these data show, including comparisons of conditional distributions.
- 24. Politics revisited.** Look again at the table of political views for the Intro Stats students in Exercise 22.
- Find the conditional distributions (percentages) of political views for the females.
 - Find the conditional distributions (percentages) of political views for the males.
 - Make a graphical display that compares the two distributions.
 - Do the variables *Politics* and *Sex* appear to be independent? Explain.

25. **Magnet schools revisited.** The *Chance* magazine article described in Exercise 9 further examined the impact of an applicant's ethnicity on the likelihood of admission to the Houston Independent School District's magnet schools programs. Those data are summarized in the table below:

		Admission Decision			Total
		Accepted	Wait-listed	Turned away	
Ethnicity	Black/Hispanic	485	0	32	517
	Asian	110	49	133	292
	White	336	251	359	946
	Total	931	300	524	1755

- What percent of all applicants were Asian?
 - What percent of the students accepted were Asian?
 - What percent of Asians were accepted?
 - What percent of all students were accepted?
26. **More politics.** Look once more at the table summarizing the political views of Intro Stats students in Exercise 22.
- Produce a graphical display comparing the conditional distributions of males and females among the three categories of politics.
 - Comment briefly on what you see from the display in a.
27. **Back to school.** Examine the table about ethnicity and acceptance for the Houston Independent School District's magnet schools program, shown in Exercise 25. Does it appear that the admissions decisions are made independent of the applicant's ethnicity? Explain.
28. **Cars.** A survey of autos parked in student and staff lots at a large university classified the brands by country of origin, as seen in the table.

		Driver	
		Student	Staff
Origin	American	107	105
	European	33	12
	Asian	55	47

- What percent of all the cars surveyed were foreign?
 - What percent of the American cars were owned by students?
 - What percent of the students owned American cars?
 - What is the marginal distribution of origin?
 - What are the conditional distributions of origin by driver classification?
 - Do you think that the origin of the car is independent of the type of driver? Explain.
29. **Weather forecasts.** Just how accurate are the weather forecasts we hear every day? The following table compares the daily forecast with a city's actual weather for a year:

		Actual Weather	
		Rain	No rain
Forecast	Rain	27	63
	No rain	7	268

- On what percent of days did it actually rain?
 - On what percent of days was rain predicted?
 - What percent of the time was the forecast correct?
 - Do you see evidence of an association between the type of weather and the ability of forecasters to make an accurate prediction? Write a brief explanation, including an appropriate graph.
30. **Twins.** In 2000, the *Journal of the American Medical Association (JAMA)* published a study that examined pregnancies that resulted in the birth of twins. Births were classified as preterm with intervention (induced labor or cesarean), preterm without procedures, or term/post-term. Researchers also classified the pregnancies by the level of prenatal medical care the mother received (inadequate, adequate, or intensive). The data, from the years 1995–1997, are summarized in the table below. Figures are in thousands of births. (*JAMA* 284 [2000]:335–341)

TWIN BIRTHS 1995–1997 (IN THOUSANDS)					
		Level of Prenatal Care			Total
		Preterm (induced or cesarean)	Preterm (without procedures)	Term or post-term	
Level of Prenatal Care	Intensive	18	15	28	61
	Adequate	46	43	65	154
	Inadequate	12	13	38	63
	Total	76	71	131	278

- What percent of these mothers received inadequate medical care during their pregnancies?
 - What percent of all twin births were preterm?
 - Among the mothers who received inadequate medical care, what percent of the twin births were preterm?
 - Create an appropriate graph comparing the outcomes of these pregnancies by the level of medical care the mother received.
 - Write a few sentences describing the association between these two variables.
31. **Blood pressure.** A company held a blood pressure screening clinic for its employees. The results are summarized in the table below by age group and blood pressure level:

		Age		
		Under 30	30–49	Over 50
Blood Pressure	Low	27	37	31
	Normal	48	91	93
	High	23	51	73

- a) Find the marginal distribution of blood pressure level.
 - b) Find the conditional distribution of blood pressure level within each age group.
 - c) Compare these distributions with a segmented bar graph.
 - d) Write a brief description of the association between age and blood pressure among these employees.
 - e) Does this prove that people’s blood pressure increases as they age? Explain.
32. **Obesity and exercise.** The Centers for Disease Control and Prevention (CDC) has estimated that 19.8% of Americans over 15 years old are obese. The CDC conducts a survey on obesity and various behaviors. Here is a table on self-reported exercise classified by body mass index (BMI):

		Body Mass Index		
		Normal (%)	Overweight (%)	Obese (%)
Physical Activity	Inactive	23.8	26.0	35.6
	Irregularly active	27.8	28.7	28.1
	Regular, not intense	31.6	31.1	27.2
	Regular, intense	16.8	14.2	9.1

- a) Are these percentages column percentages, row percentages, or table percentages?
 - b) Use graphical displays to show different percentages of physical activities for the three BMI groups.
 - c) Do these data prove that lack of exercise causes obesity? Explain.
33. **Anorexia.** Hearing anecdotal reports that some patients undergoing treatment for the eating disorder anorexia seemed to be responding positively to the antidepressant Prozac, medical researchers conducted an experiment to investigate. They found 93 women being treated for anorexia who volunteered to participate. For one year, 49 randomly selected patients were treated with Prozac and the other 44 were given an inert substance called a placebo. At the end of the year, patients were diagnosed as healthy or relapsed, as summarized in the table:

	Prozac	Placebo	Total
Healthy	35	32	67
Relapse	14	12	26
Total	49	44	93

Do these results provide evidence that Prozac might be helpful in treating anorexia? Explain.

34. **Antidepressants and bone fractures.** For a period of five years, physicians at McGill University Health Center followed more than 5000 adults over the age of 50. The

researchers were investigating whether people taking a certain class of antidepressants (SSRIs) might be at greater risk of bone fractures. Their observations are summarized in the table:

	Taking SSRI	No SSRI	Total
Experienced fractures	14	244	258
No fractures	123	4627	4750
Total	137	4871	5008

Do these results suggest there’s an association between taking SSRI antidepressants and experiencing bone fractures? Explain.

35. **Drivers’ licenses 2005.** The following table shows the number of licensed U.S. drivers by age and by sex (www.dot.gov):

Age	Male Drivers (number)	Female Drivers (number)	Total
19 and under	4,777,694	4,553,946	9,331,640
20–24	8,611,161	8,398,879	17,010,040
25–29	8,879,476	8,666,701	17,546,177
30–34	9,262,713	8,997,662	18,260,375
35–39	9,848,050	9,576,301	19,424,351
40–44	10,617,456	10,484,149	21,101,605
45–49	10,492,876	10,482,479	20,975,355
50–54	9,420,619	9,475,882	18,896,501
55–59	8,218,264	8,265,775	16,484,039
60–64	6,103,732	6,147,569	12,251,361
65–69	4,571,157	4,643,913	9,215,070
70–74	3,617,908	3,761,039	7,378,947
75–79	2,890,155	3,192,408	6,082,563
80–84	1,907,743	2,222,412	4,130,155
85 and over	1,170,817	1,406,271	2,577,088
Total	100,389,881	100,275,386	200,665,267

- a) What percent of total drivers are under 20?
 - b) What percent of total drivers are male?
 - c) Write a few sentences comparing the number of male and female licensed drivers in each age group.
 - d) Do a driver’s age and sex appear to be independent? Explain?
36. **Tattoos.** A study by the University of Texas Southwestern Medical Center examined 626 people to see if an increased risk of contracting hepatitis C was associated with having a tattoo. If the subject had a tattoo, researchers asked whether it had been done in a commercial tattoo parlor or elsewhere. Write a brief description of the association between tattooing and hepatitis C, including an appropriate graphical display.

	Tattoo done in		
	commercial parlor	Tattoo done elsewhere	No tattoo
Has hepatitis C	17	8	18
No hepatitis C	35	53	495

37. **Hospitals.** Most patients who undergo surgery make routine recoveries and are discharged as planned. Others suffer excessive bleeding, infection, or other postsurgical complications and have their discharges from the hospital delayed. Suppose your city has a large hospital and a small hospital, each performing major and minor surgeries. You collect data to see how many surgical patients have their discharges delayed by postsurgical complications, and you find the results shown in the following table.

	Discharge Delayed	
	Large hospital	Small hospital
Major surgery	120 of 800	10 of 50
Minor surgery	10 of 200	20 of 250

- Overall, for what percent of patients was discharge delayed?
 - Were the percentages different for major and minor surgery?
 - Overall, what were the discharge delay rates at each hospital?
 - What were the delay rates at each hospital for each kind of surgery?
 - The small hospital advertises that it has a lower rate of postsurgical complications. Do you agree?
 - Explain, in your own words, why this confusion occurs.
38. **Delivery service.** A company must decide which of two delivery services it will contract with. During a recent trial period, the company shipped numerous packages with each service and kept track of how often deliveries did not arrive on time. Here are the data:

Delivery Service	Type of Service	Number of Deliveries	Number of Late Packages
Pack Rats	Regular	400	12
	Overnight	100	16
Boxes R Us	Regular	100	2
	Overnight	400	28

- Compare the two services' overall percentage of late deliveries.
- On the basis of the results in part a, the company has decided to hire Pack Rats. Do you agree that Pack Rats delivers on time more often? Explain.
- The results here are an instance of what phenomenon?

39. **Graduate admissions.** A 1975 article in the magazine *Science* examined the graduate admissions process at Berkeley for evidence of sex discrimination. The table below shows the number of applicants accepted to each of four graduate programs:

Program	Males accepted (of applicants)	Females accepted (of applicants)
	1	511 of 825
2	352 of 560	17 of 25
3	137 of 407	132 of 375
4	22 of 373	24 of 341
Total	1022 of 2165	262 of 849

- What percent of total applicants were admitted?
 - Overall, was a higher percentage of males or females admitted?
 - Compare the percentage of males and females admitted in each program.
 - Which of the comparisons you made do you consider to be the most valid? Why?
40. **Be a Simpson!** Can you design a Simpson's paradox? Two companies are vying for a city's "Best Local Employer" award, to be given to the company most committed to hiring local residents. Although both employers hired 300 new people in the past year, Company A brags that it deserves the award because 70% of its new jobs went to local residents, compared to only 60% for Company B. Company B concedes that those percentages are correct, but points out that most of its new jobs were full-time, while most of Company A's were part-time. Not only that, says Company B, but a higher percentage of its full-time jobs went to local residents than did Company A's, and the same was true for part-time jobs. Thus, Company B argues, it's a better local employer than Company A.
- Show how it's possible for Company B to fill a higher percentage of both full-time and part-time jobs with local residents, even though Company A hired more local residents overall.



JUST CHECKING Answers

- 50.0%
- 44.4%
- 25.0%
- 15.6% Blue, 56.3% Brown, 28.1% Green/Hazel/Other
- 18.8% Blue, 62.5% Brown, 18.8% Green/Hazel/Other
- 40% of the blue-eyed students are female, while 50% of all students are female.
- Since blue-eyed students appear less likely to be female, it seems that *Sex* and *Eye Color* may not be independent. (But the numbers are small.)

Displaying and Summarizing Quantitative Data



Tsunamis are potentially destructive waves that can occur when the sea floor is suddenly and abruptly deformed. They are most often caused by earthquakes beneath the sea that shift the earth's crust, displacing a large mass of water.

The tsunami of December 26, 2004, with epicenter off the west coast of Sumatra, was caused by an earthquake of magnitude 9.0 on the Richter scale. It killed an estimated 297,248 people, making it the most disastrous tsunami on record. But was the earthquake that caused it truly extraordinary, or did it just happen at an unlucky place and time? The U.S. National Geophysical Data Center¹ has information on more than 2400 tsunamis dating back to 2000 B.C.E., and we have estimates of the magnitude of the underlying earthquake for 1240 of them. What can we learn from these data?

Histograms

WHO 1240 earthquakes known to have caused tsunamis for which we have data or good estimates

WHAT Magnitude (Richter scale²), depth (m), date, location, and other variables

WHEN From 2000 B.C.E. to the present

WHERE All over the earth

Let's start with a picture. For categorical variables, it is easy to draw the distribution because each category is a natural "pile." But for quantitative variables, there's no obvious way to choose piles. So, usually, we slice up all the possible values into equal-width bins. We then count the number of cases that fall into each bin. The bins, together with these counts, give the **distribution** of the quantitative variable and provide the building blocks for the histogram. By representing the counts as bars and plotting them against the bin values, the **histogram** displays the distribution at a glance.

¹ www.ngdc.noaa.gov

² Technically, Richter scale values are in units of log dyne-cm. But the Richter scale is so common now that usually the units are assumed. The U.S. Geological Survey gives the background details of Richter scale measurements on its Web site www.usgs.gov/.

For example, here are the *Magnitudes* (on the Richter scale) of the 1240 earthquakes in the NGDC data:

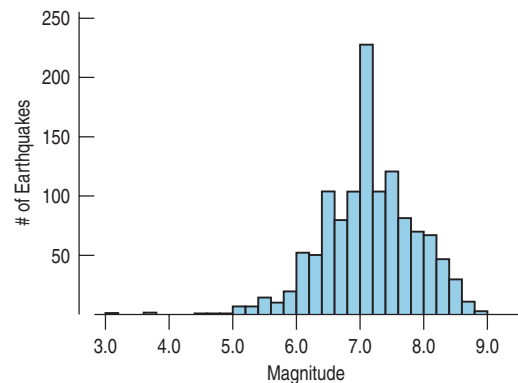


FIGURE 4.1

A histogram of earthquake magnitudes shows the number of earthquakes with magnitudes (in Richter scale units) in each bin.

One surprising feature of the earthquake magnitudes is the spike around magnitude 7.0. Only one other bin holds even half that many earthquakes. These values include historical data for which the magnitudes were estimated by experts and not measured by modern seismographs. Perhaps the experts thought 7 was a typical and reasonable value for a tsunami-causing earthquake when they lacked detailed information. That would explain the overabundance of magnitudes right at 7.0 rather than spread out near that value.

Like a bar chart, a histogram plots the bin counts as the heights of bars. In this histogram of earthquake magnitudes, each bin has a width of 0.2, so, for example, the height of the tallest bar says that there were about 230 earthquakes with magnitudes between 7.0 and 7.2. In this way, the histogram displays the entire distribution of earthquake magnitudes.

Does the distribution look as you expected? It is often a good idea to *imagine* what the distribution might look like before you make the display. That way you'll be less likely to be fooled by errors in the data or when you accidentally graph the wrong variable.

From the histogram, we can see that these earthquakes typically have magnitudes around 7. Most are between 5.5 and 8.5, and some are as small as 3 and as big as 9. Now we can answer the question about the Sumatra tsunami. With a value of 9.0 it's clear that the earthquake that caused it was an extraordinarily powerful earthquake—one of the largest on record.³

The bar charts of categorical variables we saw in Chapter 3 had spaces between the bars to separate the counts of different categories. But in a histogram, the bins slice up *all the values* of the quantitative variable, so any spaces in a histogram are actual **gaps** in the data, indicating a region where there are no values.

Sometimes it is useful to make a **relative frequency histogram**, replacing the counts on the vertical axis with the *percentage* of the total number of cases falling in each bin. Of course, the shape of the histogram is exactly the same; only the vertical scale is different.

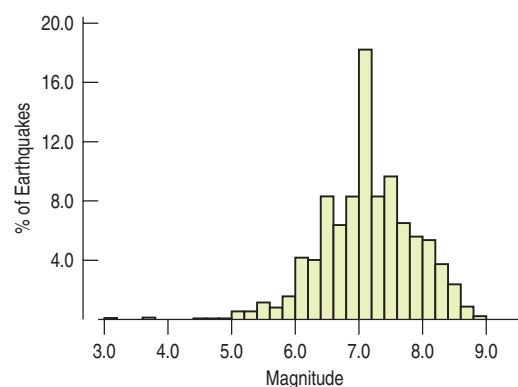


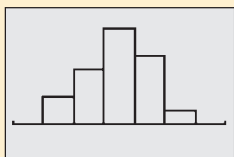
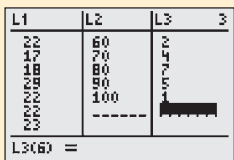
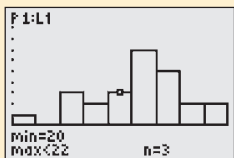
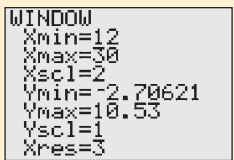
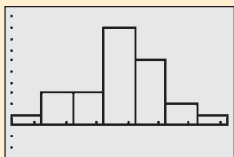
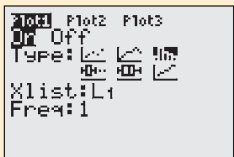
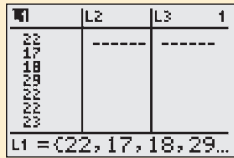
FIGURE 4.2

A relative frequency histogram looks just like a frequency histogram except for the labels on the y-axis, which now show the percentage of earthquakes in each bin.

³ Some experts now estimate the magnitude at between 9.1 and 9.3.



Making a histogram



Your calculator can create histograms. First you need some data. For an agility test, fourth-grade children jump from side to side across a set of parallel lines, counting the number of lines they clear in 30 seconds. Here are their scores:

22, 17, 18, 29, 22, 22, 23, 24, 23, 17, 21, 25, 20
12, 19, 28, 24, 22, 21, 25, 26, 25, 16, 27, 22

Enter these data into **L1**.

Now set up the calculator’s plot:

- Go to **2nd STATPLOT**, choose **Plot1**, then **ENTER**.
- In the **Plot1** screen choose **On**, select the little histogram icon, then specify **Xlist:L1** and **Freq:1**.
- Be sure to turn off any other graphs the calculator may be set up for. Just hit the **Y=** button, and deactivate any functions seen there.

All set? To create your preliminary plot go to **ZOOM**, select **9:ZoomStat**, and then **ENTER**.

You now see the calculator’s initial attempt to create a histogram of these data. Not bad. We can see that the distribution is roughly symmetric. But it’s hard to tell exactly what this histogram shows, right? Let’s fix it up a bit.

- Under **WINDOW**, let’s reset the bins to convenient, sensible values. Try **Xmin=12**, **Xmax=30** and **Xscl=2**. That specifies the range of values along the *x*-axis and makes each bar span two lines.
- Hit **GRAPH** (not **ZoomStat**—this time we want control of the scale!).

There. We still see rough symmetry, but also see that one of the scores was much lower than the others. Note that you can now find out exactly what the bars indicate by activating **TRACE** and then moving across the histogram using the arrow keys. For each bar the calculator will indicate the interval of values and the number of data values in that bin. We see that 3 kids had agility scores of 20 or 21.

Play around with the **WINDOW** settings. A different **Ymax** will make the bars appear shorter or taller. What happens if you set the bar width (**Xscl**) smaller? Or larger? You don’t want to lump lots of values into just a few bins or make so many bins that the overall shape of the histogram is not clear. Choosing the best bar width takes practice.

Finally, suppose the data are given as a frequency table. Consider a set of test scores, with two grades in the 60s, four in the 70s, seven in the 80s, five in the 90s, and one 100. Enter the group cutoffs 60, 70, 80, 90, 100 in **L2** and the corresponding frequencies 2, 4, 7, 5, 1 in **L3**. When you set up the histogram **STATPLOT**, specify **Xlist:L2** and **Freq:L3**. Can you specify the **WINDOW** settings to make this histogram look the way you want it? (By the way, if you get a **DIM MISMATCH** error, it means you can’t count. Look at **L2** and **L3**; you’ll see the two lists don’t have the same number of entries. Fix the problem by correcting the data you entered.)

Stem-and-Leaf Displays

Histograms provide an easy-to-understand summary of the distribution of a quantitative variable, but they don't show the data values themselves. Here's a histogram of the pulse rates of 24 women, taken by a researcher at a health clinic:

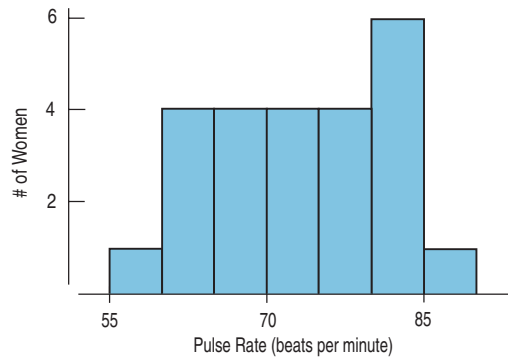


FIGURE 4.3

The pulse rates of 24 women at a health clinic

The Stem-and-Leaf display was devised by John W. Tukey, one of the greatest statisticians of the 20th century. It is called a “Stemplot” in some texts and computer programs, but we prefer Tukey’s original name for it.

The story seems pretty clear. We can see the entire span of the data and can easily see what a typical pulse rate might be. But is that all there is to these data?

A **stem-and-leaf display** is like a histogram, but it shows the individual values. It's also easier to make by hand. Here's a stem-and-leaf display of the same data:



AS **Activity: Stem-and-Leaf Displays.** As you might expect of something called “stem-and-leaf,” these displays grow as you consider each data value.

Turn the stem-and-leaf on its side (or turn your head to the right) and squint at it. It should look roughly like the histogram of the same data. Does it? Well, it's backwards because now the higher values are on the left, but other than that, it has the same shape.⁴

What does the line at the top of the display that says 8 | 8 mean? It stands for a pulse of 88 beats per minute (bpm). We've taken the tens place of the number and made that the “stem.” Then we sliced off the ones place and made it a “leaf.” The next line down is 8 | 000044. That shows that there were four pulse rates of 80 and two of 84 bpm.

Stem-and-leaf displays are especially useful when you make them by hand for batches of fewer than a few hundred data values. They are a quick way to display—and even to record—numbers. Because the leaves show the individual values, we can sometimes see even more in the data than the distribution's shape. Take another look at all the leaves of the pulse data. See anything

⁴ You could make the stem-and-leaf with the higher values on the bottom. Usually, though, higher on the top makes sense.

unusual? At a glance you can see that they are all even. With a bit more thought you can see that they are all multiples of 4—something you couldn't possibly see from a histogram. How do you think the nurse took these pulses? Counting beats for a full minute or counting for only 15 seconds and multiplying by 4?

How do stem-and-leaf displays work? Stem-and-leaf displays work like histograms, but they show more information. They use part of the number itself (called the stem) to name the bins. To make the “bars,” they use the next digit of the number. For example, if we had a test score of 83, we could write it 8|3, where 8 serves as the stem and 3 as the leaf. Then, to display the scores 83, 76, and 88 together, we would write

```

8 | 38
7 | 6
    
```

For the pulse data, we have

```

8 | 0000448
7 | 22226666
6 | 04448888
5 | 6
    Pulse Rate
    (5|6 means 56 beats/min)
    
```

This display is OK, but a little crowded. A histogram might split each line into two bars. With a stem-and-leaf, we can do the same by putting the leaves 0–4 on one line and 5–9 on another, as we saw above:

```

8 | 8
8 | 000044
7 | 6666
7 | 2222
6 | 8888
6 | 0444
5 | 6
    Pulse Rate
    (8|8 means 88 beats/min)
    
```

For numbers with three or more digits, you'll often decide to truncate (or round) the number to two places, using the first digit as the stem and the second as the leaf. So, if you had 432, 540, 571, and 638, you might display them as shown below with an indication that 6|3 means 630–639.

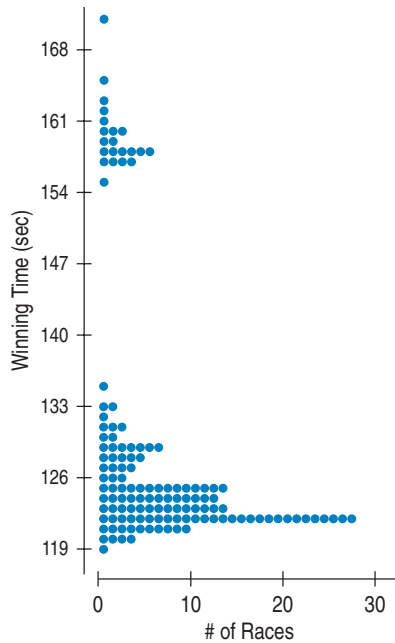
```

6 | 3
5 | 4 7
4 | 3
    
```

When you make a stem-and-leaf by hand, make sure to give each digit the same width, in order to preserve the area principle. (That can lead to some fat 1's and thin 8's—but it makes the display honest.)

Dotplots

AS **Activity: Dotplots.** Click on points to see their values and even drag them around.



A **dotplot** is a simple display. It just places a dot along an axis for each case in the data. It's like a stem-and-leaf display, but with dots instead of digits for all the leaves. Dotplots are a great way to display a small data set (especially if you forget how to write the digits from 0 to 9). Here's a dotplot of the time (in seconds) that the winning horse took to win the Kentucky Derby in each race between the first Derby in 1875 and the 2008 Derby.

Dotplots show basic facts about the distribution. We can find the slowest and quickest races by finding times for the topmost and bottommost dots. It's also clear that there are two clusters of points, one just below 160 seconds and the other at about 122 seconds. Something strange happened to the Derby times. Once we know to look for it, we can find out that in 1896 the distance of the Derby race was changed from 1.5 miles to the current 1.25 miles. That explains the two clusters of winning times.

Some dotplots stretch out horizontally, with the counts on the vertical axis, like a histogram. Others, such as the one shown here, run vertically, like a stem-and-leaf display. Some dotplots place points next to each other when they would otherwise overlap. Others just place them on top of one another. Newspapers sometimes offer dotplots with the dots made up of little pictures.

FIGURE 4.4

A dotplot of Kentucky Derby winning times plots each race as its own dot, showing the bimodal distribution.

Think Before You Draw, Again

Suddenly, we face a lot more options when it's time to invoke our first rule of data analysis and make a picture. You'll need to *Think* carefully to decide which type of graph to make. In the previous chapter you learned to check the Categorical Data Condition before making a pie chart or a bar chart. Now, before making a stem-and-leaf display, a histogram, or a dotplot, you need to check the

Quantitative Data Condition: The data are values of a quantitative variable whose units are known.

Although a bar chart and a histogram may look somewhat similar, they're not the same display. You can't display categorical data in a histogram or quantitative data in a bar chart. Always check the condition that confirms what type of data you have before proceeding with your display.

Step back from a histogram or stem-and-leaf display. What can you say about the distribution? When you describe a distribution, you should always tell about three things: its **shape**, **center**, and **spread**.

The Shape of a Distribution

1. Does the histogram have a single, central hump or several separated humps? These humps are called **modes**.⁵ The earthquake magnitudes have a single mode

⁵ Well, technically, it's the value on the horizontal axis of the histogram that is the mode, but anyone asked to point to the mode would point to the hump.

The **mode** is sometimes defined as the single value that appears most often. That definition is fine for categorical variables because all we need to do is count the number of cases for each category. For quantitative variables, the mode is more ambiguous. What is the mode of the Kentucky Derby times? Well, seven races were timed at 122.2 seconds—more than any other race time. Should that be the mode? Probably not. For quantitative data, it makes more sense to use the term “mode” in the more general sense of the peak of the histogram rather than as a single summary value. In this sense, the important feature of the Kentucky Derby races is that there are two distinct modes, representing the two different versions of the race and warning us to consider those two versions separately.

at just about 7. A histogram with one peak, such as the earthquake magnitudes, is dubbed **unimodal**; histograms with two peaks are **bimodal**, and those with three or more are called **multimodal**.⁶ For example, here’s a bimodal histogram.

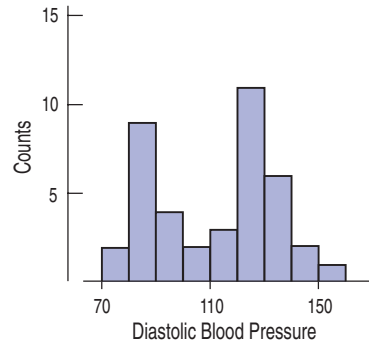


FIGURE 4.5
A bimodal histogram has two apparent peaks.

A histogram that doesn’t appear to have any mode and in which all the bars are approximately the same height is called **uniform**.

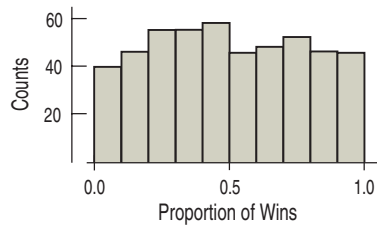
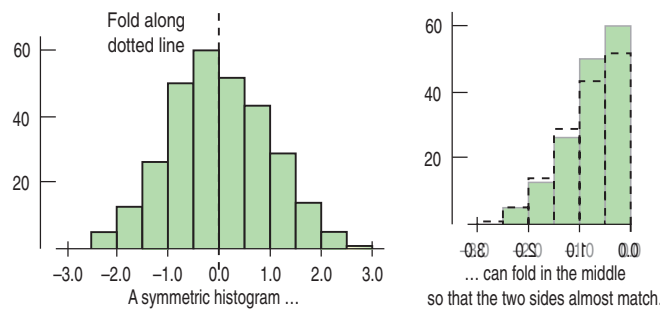


FIGURE 4.6
In a uniform histogram, the bars are all about the same height. The histogram doesn’t appear to have a mode.

You’ve heard of pie à la mode. Is there a connection between pie and the mode of a distribution? Actually, there is! The mode of a distribution is a *popular* value near which a lot of the data values gather. And “à la mode” means “in style”—not “with ice cream.” That just happened to be a *popular* way to have pie in Paris around 1900.

2. *Is the histogram symmetric?* Can you fold it along a vertical line through the middle and have the edges match pretty closely, or are more of the values on one side?



The (usually) thinner ends of a distribution are called the **tails**. If one tail stretches out farther than the other, the histogram is said to be **skewed** to the side of the longer tail.

⁶ Apparently, statisticians don’t like to count past two.

AS **Activity: Attributes of Distribution Shape.** This activity and the others on this page show off aspects of distribution shape through animation and example, then let you make and interpret histograms with your statistics package.

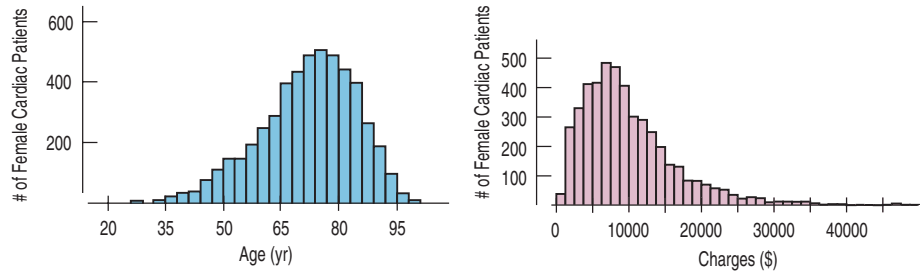
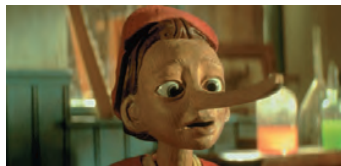


FIGURE 4.8
Two skewed histograms showing data on two variables for all female heart attack patients in New York state in one year. The blue one (age in years) is skewed to the left. The purple one (charges in \$) is skewed to the right.



3. Do any unusual features stick out? Often such features tell us something interesting or exciting about the data. You should always mention any stragglers, or **outliers**, that stand off away from the body of the distribution. If you're collecting data on nose lengths and Pinocchio is in the group, you'd probably notice him, and you'd certainly want to mention it.

Outliers can affect almost every method we discuss in this course. So we'll always be on the lookout for them. An outlier can be the most informative part of your data. Or it might just be an error. But don't throw it away without comment. Treat it specially and discuss it when you tell about your data. Or find the error and fix it if you can. Be sure to look for outliers. Always.

In the next chapter you'll learn a handy rule of thumb for deciding when a point might be considered an outlier.

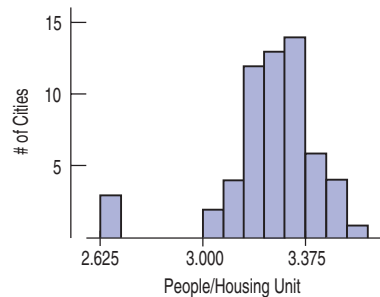


FIGURE 4.9
A histogram with outliers. There are three cities in the leftmost bar.

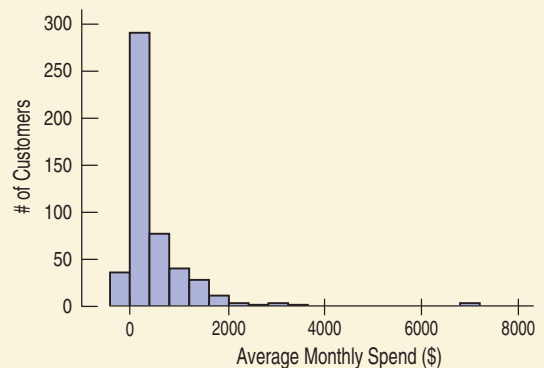
FOR EXAMPLE

Describing histograms

A credit card company wants to see how much customers in a particular segment of their market use their credit card. They have provided you with data⁷ on the amount spent by 500 selected customers during a 3-month period and have asked you to summarize the expenditures. Of course, you begin by making a histogram.

Question: Describe the shape of this distribution.

The distribution of expenditures is unimodal and skewed to the high end. There is an extraordinarily large value at about \$7000, and some of the expenditures are negative.



⁷These data are real, but cannot be further identified for obvious privacy reasons.

Are there any gaps in the distribution? The Kentucky Derby data that we saw in the dotplot on page 49 has a large gap between two groups of times, one near 120 seconds and one near 160. Gaps help us see multiple modes and encourage us to notice when the data may come from different sources or contain more than one group.



Toto, I've a feeling we're not in math class anymore . . . When Dorothy and her dog Toto land in Oz, everything is more vivid and colorful, but also more dangerous and exciting. Dorothy has new choices to make. She can't always rely on the old definitions, and the yellow brick road has many branches. You may be coming to a similar realization about Statistics.

When we summarize data, our goal is usually more than just developing a detailed knowledge of the data we have at hand. Scientists generally don't care about the particular guinea pigs they've treated, but rather about what their reactions say about how animals (and, perhaps, humans) would respond.

When you look at data, you want to know what the data say about the world, so you'd like to know whether the patterns you see in histograms and summary statistics generalize to other individuals and situations. You'll want to calculate summary statistics accurately, but then you'll also want to think about what they may say beyond just describing the data. And your knowledge about the world matters when you think about the overall meaning of your analysis.

It may surprise you that many of the most important concepts in Statistics are not defined as precisely as most concepts in mathematics. That's done on purpose, to leave room for judgment.

Because we want to see broader patterns rather than focus on the details of the data set we're looking at, we deliberately leave some statistical concepts a bit vague. Whether a histogram is symmetric or skewed, whether it has one or more modes, whether a point is far enough from the rest of the data to be considered an outlier—these are all somewhat vague concepts. And they all require judgment. You may be used to finding a single correct and precise answer, but in Statistics, there may be more than one interpretation. That may make you a little uncomfortable at first, but soon you'll see that this room for judgment brings you enormous power and responsibility. It means that using your own knowledge and judgment and supporting your findings with statistical evidence and justifications entitles you to your own opinions about what you see.



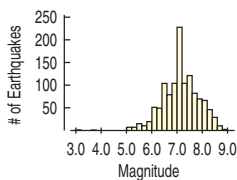
JUST CHECKING

It's often a good idea to think about what the distribution of a data set might look like before we collect the data. What do you think the distribution of each of the following data sets will look like? Be sure to discuss its shape. Where do you think the center might be? How spread out do you think the values will be?

1. Number of miles run by Saturday morning joggers at a park.
2. Hours spent by U.S. adults watching football on Thanksgiving Day.
3. Amount of winnings of all people playing a particular state's lottery last week.
4. Ages of the faculty members at your school.
5. Last digit of phone numbers on your campus.

The Center of the Distribution: The Median

Let's return to the tsunami earthquakes. But this time, let's look at just 25 years of data: 176 earthquakes that occurred from 1981 through 2005. These should be more accurately measured than prehistoric quakes because seismographs were in wide use. Try to put your finger on the histogram at the value you think is



typical. (Read the value from the horizontal axis and remember it.) When we think of a typical value, we usually look for the center of the distribution. Where do you think the center of this distribution is? For a unimodal, symmetric distribution such as these earthquake data, it's easy. We'd all agree on the center of symmetry, where we would fold the histogram to match the two sides. But when the distribution is skewed or possibly multimodal, it's not immediately clear what we even mean by the center.

One reasonable choice of typical value is the value that is literally in the middle, with half the values below it and half above it.

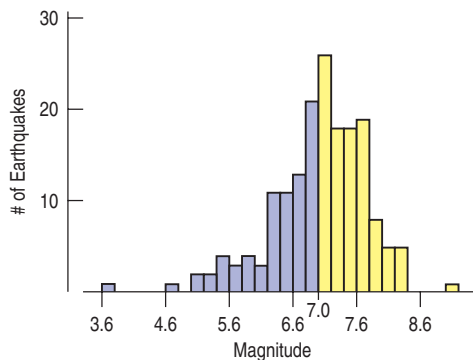


FIGURE 4.10 *Tsunami-causing earthquakes (1981–2005)*

The median splits the histogram into two halves of equal area.

Histograms follow the area principle, and each half of the data has about 88 earthquakes, so each colored region has the same area in the display. The middle value that divides the histogram into two equal areas is called the **median**.

The median has the same units as the data. Be sure to include the units whenever you discuss the median.

For the recent tsunamis, there are 176 earthquakes, so the median is found at the $(176 + 1)/2 = 88.5$ th place in the sorted data. That “.5” just says to average the two values on either side: the 88th and the 89th. The median earthquake magnitude is 7.0.

NOTATION ALERT:

We always use n to indicate the number of values. Some people even say, “How big is the n ?” when they mean the number of data values.

How do medians work? Finding the median of a batch of n numbers is easy as long as you remember to order the values first. If n is odd, the median is the middle value. Counting in from the ends, we find this value in the $\frac{n + 1}{2}$ position.

When n is even, there are two middle values. So, in this case, the median is the average of the two values in positions $\frac{n}{2}$ and $\frac{n}{2} + 1$.

Here are two examples:

Suppose the batch has these values: 14.1, 3.2, 25.3, 2.8, -17.5, 13.9, 45.8. First we order the values: -17.5, 2.8, 3.2, 13.9, 14.1, 25.3, 45.8.

Since there are 7 values, the median is the $(7 + 1)/2 = 4$ th value, counting from the top or bottom: 13.9. Notice that 3 values are lower, 3 higher.

Suppose we had the same batch with another value at 35.7. Then the ordered values are -17.5, 2.8, 3.2, 13.9, 14.1, 25.3, 35.7, 45.8.

The median is the average of the $8/2$ or 4th, and the $(8/2) + 1$, or 5th, values. So the median is $(13.9 + 14.1)/2 = 14.0$. Four data values are lower, and four higher.

The median is one way to find the center of the data. But there are many others. We'll look at an even more important measure later in this chapter.

Knowing the median, we could say that a typical tsunami-causing earthquake, worldwide, was about 7.0 on the Richter scale. How much does that really say? How well does the median describe the data? After all, not every earthquake has a Richter scale value of 7.0. Whenever we find the center of data, the next step is always to ask how well it actually summarizes the data.

Spread: Home on the Range

Statistics pays close attention to what we *don't* know as well as what we do know. Understanding how spread out the data are is a first step in understanding what a summary *cannot* tell us about the data. It's the beginning of telling us what we don't know.

If every earthquake that caused a tsunami registered 7.0 on the Richter scale, then knowing the median would tell us everything about the distribution of earthquake magnitudes. The more the data vary, however, the less the median alone can tell us. So we need to measure how much the data values vary around the center. In other words, how spread out are they? When we describe a distribution numerically, we always report a measure of its **spread** along with its center.

How should we measure the spread? We could simply look at the extent of the data. How far apart are the two extremes? The **range** of the data is defined as the *difference* between the maximum and minimum values:

$$\text{Range} = \text{max} - \text{min}.$$

Notice that the range is a *single number*, not an interval of values, as you might think from its use in common speech. The maximum magnitude of these earthquakes is 9.0 and the minimum is 3.7, so the *range* is $9.0 - 3.7 = 5.3$.

The range has the disadvantage that a single extreme value can make it very large, giving a value that doesn't really represent the data overall.

Spread: The Interquartile Range

A better way to describe the spread of a variable might be to ignore the extremes and concentrate on the middle of the data. We could, for example, find the range of just the middle half of the data. What do we mean by the middle half? Divide the data in half at the median. Now divide both halves in half again, cutting the data into four quarters. We call these new dividing points **quartiles**. One quarter of the data lies below the **lower quartile**, and one quarter of the data lies above the **upper quartile**, so half the data lies between them. The quartiles border the middle half of the data.

How do quartiles work? A simple way to find the quartiles is to start by splitting the batch into two halves at the median. (When n is odd, some statisticians include the median in both halves; others omit it.) The lower quartile is the median of the lower half, and the upper quartile is the median of the upper half.

Here are our two examples again.

The ordered values of the first batch were $-17.5, 2.8, 3.2, 13.9, 14.1, 25.3,$ and 45.8 , with a median of 13.9 . Excluding the median, the two halves of the list are $-17.5, 2.8, 3.2$ and $14.1, 25.3, 45.8$.

Each half has 3 values, so the median of each is the middle one. The lower quartile is 2.8 , and the upper quartile is 25.3 .

The second batch of data had the ordered values $-17.5, 2.8, 3.2, 13.9, 14.1, 25.3, 35.7,$ and 45.8 .

Here n is even, so the two halves of 4 values are $-17.5, 2.8, 3.2, 13.9$ and $14.1, 25.3, 35.7, 45.8$.

Now the lower quartile is $(2.8 + 3.2)/2 = 3.0$, and the upper quartile is $(25.3 + 35.7)/2 = 30.5$.

The difference between the quartiles tells us how much territory the middle half of the data covers and is called the **interquartile range**. It's commonly abbreviated **IQR** (and pronounced "eye-cue-are," not "ikker"):

$$IQR = \text{upper quartile} - \text{lower quartile}.$$

For the earthquakes, there are 88 values below the median and 88 values above the median. The midpoint of the lower half is the average of the 44th and 45th values in the ordered data; that turns out to be 6.6. In the upper half we average the 132nd and 133rd values, finding a magnitude of 7.6 as the third quartile. The *difference* between the quartiles gives the IQR:

$$IQR = 7.6 - 6.6 = 1.0.$$

Now we know that the middle half of the earthquake magnitudes extends across a (interquartile) range of 1.0 Richter scale units. This seems like a reasonable summary of the spread of the distribution, as we can see from this histogram:

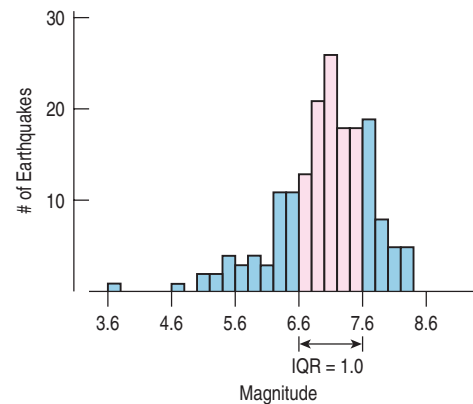


FIGURE 4.11
The quartiles bound the middle 50% of the values of the distribution. This gives a visual indication of the spread of the data. Here we see that the IQR is 1.0 Richter scale units.

The IQR is almost always a reasonable summary of the spread of a distribution. Even if the distribution itself is skewed or has some outliers, the IQR should provide useful information. The one exception is when the data are strongly bimodal. For example, remember the dotplot of winning times in the Kentucky Derby (page 49)? Because the race distance was changed, we really have data on two different races, and they shouldn't be summarized together.

So, what is a quartile anyway? Finding the quartiles sounds easy, but surprisingly, the quartiles are not well-defined. It's not always clear how to find a value such that exactly one quarter of the data lies above or below that value. We offered a simple rule for Finding Quartiles in the box on page 54: Find the median of each half of the data split by the median. When n is odd, we (and your TI calculator) omit the median from each of the halves. Some other texts include the median in both halves before finding the quartiles. Both methods are commonly used. If you are willing to do a bit more calculating, there are several other methods that locate a quartile somewhere between adjacent data values. We know of at least six different rules for finding quartiles. Remarkably, each one is in use in some software package or calculator.

So don't worry too much about getting the "exact" value for a quartile. All of the methods agree pretty closely when the data set is large. When the data set is small, different rules will disagree more, but in that case there's little need to summarize the data anyway.

Remember, Statistics is about understanding the world, not about calculating the right number. The "answer" to a statistical question is a sentence about the issue raised in the question.

The lower and upper quartiles are also known as the 25th and 75th percentiles of the data, respectively, since the lower quartile falls above 25% of the data and the upper quartile falls above 75% of the data. If we count this way, the median is the 50th percentile. We could, of course, define and calculate any percentile that we want. For example, the 10th percentile would be the number that falls above the lowest 10% of the data values.

5-Number Summary

NOTATION ALERT:

We always use Q1 to label the lower (25%) quartile and Q3 to label the upper (75%) quartile. We skip the number 2 because the median would, by this system, naturally be labeled Q2—but we don't usually call it that.

The **5-number summary** of a distribution reports its median, quartiles, and extremes (maximum and minimum). The 5-number summary for the recent tsunami earthquake *Magnitudes* looks like this:

Max	9.0
Q3	7.6
Median	7.0
Q1	6.6
Min	3.7

It's good idea to report the number of data values and the identity of the cases (the *Who*). Here there are 176 earthquakes.

The 5-number summary provides a good overview of the distribution of magnitudes of these tsunami-causing earthquakes. For a start, we can see that the median magnitude is 7.0. Because the IQR is only $7.6 - 6.6 = 1$, we see that many quakes are close to the median magnitude. Indeed, the quartiles show us that the middle half of these earthquakes had magnitudes between 6.6 and 7.6. One quarter of the earthquakes had magnitudes above 7.6, although one tsunami was caused by a quake measuring only 3.7 on the Richter scale.

STEP-BY-STEP EXAMPLE

Shape, Center, and Spread: Flight Cancellations



The U.S. Bureau of Transportation Statistics (www.bts.gov) reports data on airline flights. Let's look at data giving the percentage of flights cancelled each month between 1995 and 2005.

Question: How often are flights cancelled?

WHO	Months
WHAT	Percentage of flights cancelled at U.S. airports
WHEN	1995–2005
WHERE	United States



Variable: Identify the *variable*, and decide how you wish to display it.

To identify a variable, report the W's.

Select an appropriate display based on the nature of the data and what you want to know.

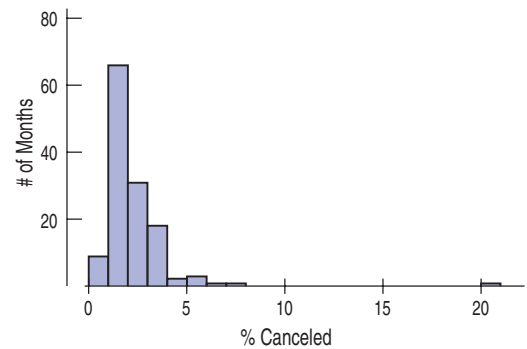
I want to learn about the monthly percentage of flight cancellations at U.S. airports.

I have data from the U.S. Bureau of Transportation Statistics giving the percentage of flights cancelled at U.S. airports each month between 1995 and 2005.

✓ **Quantitative Data Condition:** Percentages are quantitative. A histogram and numerical summaries would be appropriate.



Mechanics: We usually make histograms with a computer or graphing calculator.



The histogram shows a distribution skewed to the high end and one extreme outlier, a month in which more than 20% of flights were cancelled.

In most months, fewer than 5% of flights are cancelled and usually only about 2% or 3%. That seems reasonable.



It's always a good idea to think about what you expect to see so that you can check whether the histogram looks like what you expected.

With 132 cases, we probably have more data than you'd choose to work with by hand. The results given here are from technology.

Count	132
Max	20.240
Q3	2.615
Median	1.755
Q1	1.445
Min	0.770
IQR	1.170

TELL

Interpretation: Describe the shape, center, and spread of the distribution. Report on the symmetry, number of modes, and any gaps or outliers. You should also mention any concerns you may have about the data.

The distribution of cancellations is skewed to the right, and this makes sense: The values can't fall below 0%, but can increase almost arbitrarily due to bad weather or other events.

The median is 1.76% and the IQR is 1.17%. The low IQR indicates that in most months the cancellation rate is close to the median. In fact, it's between 1.4% and 2.6% in the middle 50% of all months, and in only 1/4 of the months were more than 2.6% of flights cancelled.

There is one extraordinary value: 20.2%. Looking it up, I find that the extraordinary month was September 2001. The attacks of September 11 shut down air travel for several days, accounting for this outlier.

Summarizing Symmetric Distributions: The Mean

NOTATION ALERT:

In Algebra you used letters to represent values in a problem, but it didn't matter what letter you picked. You could call the width of a rectangle X or you could call it w (or *Fred*, for that matter). But in Statistics, the notation is part of the vocabulary. For example, in Statistics n is always the number of data values. Always.

We have already begun to point out such special notation conventions: n , $Q1$, and $Q3$. Think of them as part of the terminology you need to learn in this course.

Here's another one: Whenever we put a bar over a symbol, it means "find the mean."

Medians do a good job of summarizing the center of a distribution, even when the shape is skewed or when there is an outlier, as with the flight cancellations. But when we have symmetric data, there's another alternative. You probably already know how to average values. In fact, to find the median when n is even, we said you should average the two middle values, and you didn't even flinch.

The earthquake magnitudes are pretty close to symmetric, so we can also summarize their center with a mean. The mean tsunami earthquake magnitude is 6.96—about what we might expect from the histogram. You already know how to average values, but this is a good place to introduce notation that we'll use throughout the book. We use the Greek capital letter sigma, Σ , to mean "sum" (sigma is "S" in Greek), and we'll write:

$$\bar{y} = \frac{\text{Total}}{n} = \frac{\sum y}{n}.$$

The formula says to add up all the values of the variable and divide that sum by the number of data values, n —just as you've always done.⁸

Once we've averaged the data, you'd expect the result to be called the *average*, but that would be too easy. Informally, we speak of the "average person" but we don't add up people and divide by the number of people. A median is also a kind of average. To make this distinction, the value we calculated is called the mean, \bar{y} , and pronounced "y-bar."

⁸ You may also see the variable called x and the equation written $\bar{x} = \frac{\text{Total}}{n} = \frac{\sum x}{n}$. Don't let that throw you. You are free to name the variable anything you want, but we'll generally use y for variables like this that we want to summarize, model, or predict. (Later we'll talk about variables that are used to explain, model, or predict y . We'll call them x .)

In everyday language, sometimes “average” does mean what we want it to mean. We don’t talk about your grade point mean or a baseball player’s batting mean or the Dow Jones Industrial mean. So we’ll continue to say “average” when that seems most natural. When we do, though, you may assume that what we mean is the mean.

The **mean** feels like the center because it is the point where the histogram balances:

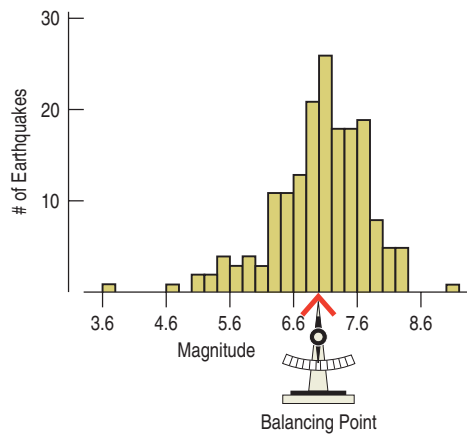


FIGURE 4.12
The mean is located at the balancing point of the histogram.

Mean or Median?

Using the center of balance makes sense when the data are symmetric. But data are not always this well behaved. If the distribution is skewed or has outliers, the center is not so well defined and the mean may not be what we want. For example, the mean of the flight cancellations doesn’t give a very good idea of the typical percentage of cancellations.

TI-nspire
Mean, median, and outliers.
Drag data points around to explore how outliers affect the mean and median.

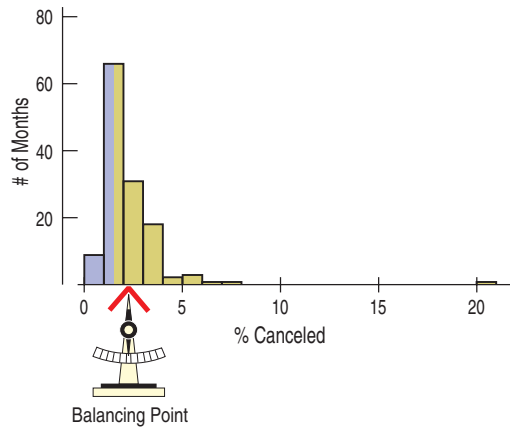


FIGURE 4.13
The median splits the area of the histogram in half at 1.75%. Because the distribution is skewed to the right, the mean (2.28%) is higher than the median. The points at the right have pulled the mean toward them away from the median.

A S **Activity: The Center of a Distribution.** Compare measures of center by dragging points up and down and seeing the consequences. Another activity shows how to find summaries with your statistics package.

The mean is 2.28%, but nearly 70% of months had cancellation rates below that, so the mean doesn’t feel like a good overall summary. Why is the balancing point so high? The large outlying value pulls it to the right. For data like these, the median is a better summary of the center.

Because the median considers only the order of the values, it is **resistant to values that are extraordinarily large or small**; it simply notes that they are one of the “big ones” or the “small ones” and ignores their distance from the center.

For the tsunami earthquake magnitudes, it doesn’t seem to make much difference—the mean is 6.9; the median is 7.0. When the data are symmetric, the mean and median will be close, but when the data are skewed, the median is likely to be a better choice. So, why not just use the median? Well, for one, the median can go overboard. It’s not just resistant to occasional outliers, but can be unaffected by changes in up to half the data values. By contrast, the mean includes input from

each data value and gives each one equal weight. It's also easier to work with, so when the distribution is unimodal and symmetric, we'll use the mean.

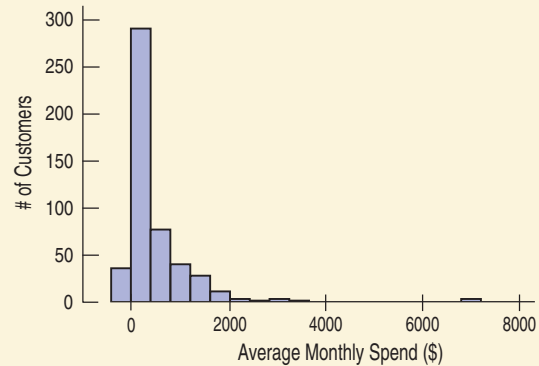
Of course, to choose between mean and median, we'll start by looking at the data. If the histogram is symmetric and there are no outliers, we'll prefer the mean. However, if the histogram is skewed or has outliers, we're usually better off with the median. If you're not sure, report both and discuss why they might differ.

FOR EXAMPLE

Describing center

Recap: You want to summarize the expenditures of 500 credit card company customers, and have looked at a histogram.

Question: You have found the mean expenditure to be \$478.19 and the median to be \$216.28. Which is the more appropriate measure of center, and why?



Because the distribution of expenditures is skewed, the median is the more appropriate measure of center. Unlike the mean, it's not affected by the large outlying value or by the skewness. Half of these credit card customers had average monthly expenditures less than \$216.28 and half more.

When to expect skewness Even without making a histogram, we can expect some variables to be skewed. When values of a quantitative variable are bounded on one side but not the other, the distribution may be skewed. For example, incomes and waiting times can't be less than zero, so they are often skewed to the right. Amounts of things (dollars, employees) are often skewed to the right for the same reason. If a test is too easy, the distribution will be skewed to the left because many scores will bump against 100%. And combinations of things are often skewed. In the case of the cancelled flights, flights are more likely to be cancelled in January (due to snowstorms) and in August (thunderstorms). Combining values across months leads to a skewed distribution.

What About Spread? The Standard Deviation

AS **Activity: The Spread of a Distribution.** What happens to measures of spread when data values change may not be quite what you expect.

The IQR is always a reasonable summary of spread, but because it uses only the two quartiles of the data, it ignores much of the information about how individual values vary. A more powerful approach uses the **standard deviation**, which takes into account how far *each* value is from the mean. Like the mean, the standard deviation is appropriate only for symmetric data.

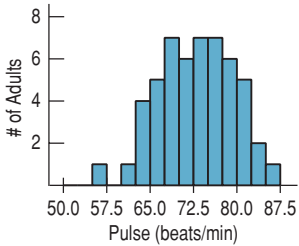
One way to think about spread is to examine how far each data value is from the mean. This difference is called a *deviation*. We could just average the deviations, but the positive and negative differences always cancel each other out. So the average deviation is always zero—not very helpful.

To keep them from canceling out, we *square* each deviation. Squaring always gives a positive value, so the sum won't be zero. That's great. Squaring also emphasizes larger differences—a feature that turns out to be both good and bad.

NOTATION ALERT:

s^2 always means the variance of a set of data, and s always denotes the standard deviation.

WHO 52 adults
WHAT Resting heart rates
UNITS Beats per minute



When we add up these squared deviations and find their average (almost), we call the result the **variance**:

$$s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}$$

Why almost? It *would* be a mean if we divided the sum by n . Instead, we divide by $n - 1$. Why? The simplest explanation is “to drive you crazy.” But there are good technical reasons, some of which we’ll see later.

The variance will play an important role later in this book, but it has a problem as a measure of spread. Whatever the units of the original data are, the variance is in *squared* units. We want measures of spread to have the same units as the data. And we probably don’t want to talk about squared dollars or *mpg*². So, to get back to the original units, we take the square root of s^2 . The result, s , is the **standard deviation**.

Putting it all together, the standard deviation of the data is found by the following formula:

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$$

You will almost always rely on a calculator or computer to do the calculating.

Understanding what the standard deviation really means will take some time, and we’ll revisit the concept in later chapters. For now, have a look at this histogram of resting pulse rates. The distribution is roughly symmetric, so it’s okay to choose the mean and standard deviation as our summaries of center and spread. The mean pulse rate is 72.7 beats per minute, and we can see that’s a typical heart rate. We also see that some heart rates are higher and some lower—but how much? Well, the standard deviation of 6.5 beats per minute indicates that, on average, we might expect people’s heart rates to differ from the mean rate by about 6.5 beats per minute. Looking at the histogram, we can see that 6.5 beats above or below the mean appears to be a typical deviation.

How does standard deviation work? To find the standard deviation, start with the mean, \bar{y} . Then find the *deviations* by taking \bar{y} from each value: $(y - \bar{y})$. Square each deviation: $(y - \bar{y})^2$.

Now you’re nearly home. Just add these up and divide by $n - 1$. That gives you the variance, s^2 . To find the standard deviation, s , take the square root. Here we go: Suppose the batch of values is 14, 13, 20, 22, 18, 19, and 13.

The mean is $\bar{y} = 17$. So the deviations are found by subtracting 17 from each value:

Original Values	Deviations	Squared Deviations
14	$14 - 17 = -3$	$(-3)^2 = 9$
13	$13 - 17 = -4$	$(-4)^2 = 16$
20	$20 - 17 = 3$	9
22	$22 - 17 = 5$	25
18	$18 - 17 = 1$	1
19	$19 - 17 = 2$	4
13	$13 - 17 = -4$	16

Add up the squared deviations: $9 + 16 + 9 + 25 + 1 + 4 + 16 = 80$.

Now divide by $n - 1$: $80/6 = 13.33$.

Finally, take the square root: $s = \sqrt{13.33} = 3.65$

Thinking About Variation

A S

Activity: Displaying

Spread. What does the standard deviation look like on a histogram? How about the IQR?

Why do banks favor a single line that feeds several teller windows rather than separate lines for each teller? The average waiting time is the same. But the time you can expect to wait is less variable when there is a single line, and people prefer consistency.

Statistics is about variation, so spread is an important fundamental concept in Statistics. Measures of spread help us to be precise about what we *don't* know. If many data values are scattered far from the center, the IQR and the standard deviation will be large. If the data values are close to the center, then these measures of spread will be small. If all our data values were exactly the same, we'd have no question about summarizing the center, and all measures of spread would be zero—and we wouldn't need Statistics. You might think this would be a big plus, but it would make for a boring world. Fortunately (at least for Statistics), data do vary.

Measures of spread tell how well other summaries describe the data. That's why we always (always!) report a spread along with any summary of the center.



JUST CHECKING

6. The U.S. Census Bureau reports the median family income in its summary of census data. Why do you suppose they use the median instead of the mean? What might be the disadvantages of reporting the mean?
7. You've just bought a new car that claims to get a highway fuel efficiency of 31 miles per gallon. Of course, your mileage will "vary." If you had to guess, would you expect the IQR of gas mileage attained by all cars like yours to be 30 mpg, 3 mpg, or 0.3 mpg? Why?
8. A company selling a new MP3 player advertises that the player has a mean lifetime of 5 years. If you were in charge of quality control at the factory, would you prefer that the standard deviation of lifespans of the players you produce be 2 years or 2 months? Why?

What to Tell About a Quantitative Variable

A S

Activity: Playing with

Summaries. Here's a Statistics game about summaries that even some experienced statisticians find . . . well, challenging. Your intuition may be better. Give it a try!

Ti-*nspire*

Standard deviation, IQR, and outliers. Drag data points around to explore how outliers affect measures of spread.


What should you *Tell* about a quantitative variable?

- ▶ Start by making a histogram or stem-and-leaf display, and discuss the shape of the distribution.
- ▶ Next, discuss the center *and* spread.
 - ▶ We always pair the median with the IQR and the mean with the standard deviation. It's not useful to report one without the other. Reporting a center without a spread is dangerous. You may think you know more than you do about the distribution. Reporting only the spread leaves us wondering where we are.
 - ▶ If the shape is skewed, report the median and IQR. You may want to include the mean and standard deviation as well, but you should point out why the mean and median differ.
 - ▶ If the shape is symmetric, report the mean and standard deviation and possibly the median and IQR as well. For unimodal symmetric data, the IQR is usually a bit larger than the standard deviation. If that's not true of your data set, look again to make sure that the distribution isn't skewed and there are no outliers.

How “Accurate” Should We Be?

Don’t think you should report means and standard deviations to a zillion decimal places; such implied accuracy is really meaningless. Although there is no ironclad rule, statisticians commonly report summary statistics to one or two decimal places more than the original data have.

- ▶ Also, discuss any unusual features.
 - ▶ If there are multiple modes, try to understand why. If you can identify a reason for separate modes (for example, women and men typically have heart attacks at different ages), it may be a good idea to split the data into separate groups.
 - ▶ If there are any clear outliers, you should point them out. If you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be revealing. (Of course, the median and IQR won’t be affected very much by the outliers.)

STEP-BY-STEP EXAMPLE		Summarizing a distribution
<p>One of the authors owned a 1989 Nissan Maxima for 8 years. Being a statistician, he recorded the car’s fuel efficiency (in mpg) each time he filled the tank. He wanted to know what fuel efficiency to expect as “ordinary” for his car. (Hey, he’s a statistician. What would you expect?⁹) Knowing this, he was able to predict when he’d need to fill the tank again and to notice if the fuel efficiency suddenly got worse, which could be a sign of trouble.</p> <p>Question: How would you describe the distribution of <i>Fuel efficiency</i> for this car?</p>		
	<p>Plan State what you want to find out.</p> <p>Variable Identify the variable and report the W’s.</p> <p>Be sure to check the appropriate condition.</p>	<p>I want to summarize the distribution of Nissan Maxima fuel efficiency.</p> <p>The data are the fuel efficiency values in miles per gallon for the first 100 fill-ups of a 1989 Nissan Maxima between 1989 and 1992.</p> <p>✓ Quantitative Data Condition: The fuel efficiencies are quantitative with units of miles per gallon. Histograms and boxplots are appropriate displays for displaying the distribution. Numerical summaries are appropriate as well.</p>

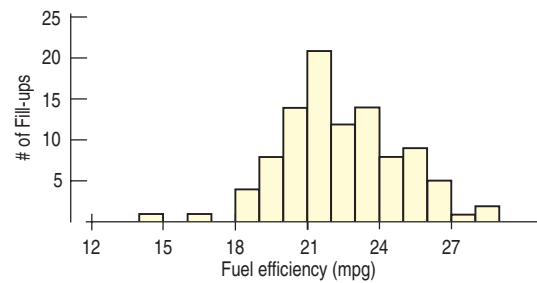
⁹ He also recorded the time of day, temperature, price of gas, and phase of the moon. (OK, maybe not phase of the moon.) His data are on the DVD.

SHOW

Mechanics Make a histogram and boxplot. Based on the shape, choose appropriate numerical summaries.

REALITY CHECK

A value of 22 mpg seems reasonable for such a car. The spread is reasonable, although the range looks a bit large.



A histogram of the data shows a fairly symmetric distribution with a low outlier.

Count	100
Mean	22.4 mpg
StdDev	2.45
Q1	20.8
Median	22.0
Q3	24.0
IQR	3.2

The mean and median are close, so the outlier doesn't seem to be a problem. I can use the mean and standard deviation.

TELL

Conclusion Summarize and interpret your findings in context. Be sure to discuss the distribution's shape, center, spread, and unusual features (if any).

The distribution of mileage is unimodal and roughly symmetric with a mean of 22.4 mpg. There is a low outlier that should be investigated, but it does not influence the mean very much. The standard deviation suggests that from tankful to tankful, I can expect the car's fuel economy to differ from the mean by an average of about 2.45 mpg.

Are my statistics "right"? When you calculate a mean, the computation is clear: You sum all the values and divide by the sample size. You may round your answer less or more than someone else (we recommend one more decimal place than the data), but all books and technologies agree on how to find the mean. Some statistics, however, are more problematic. For example we've already pointed out that methods of finding quartiles differ.

Differences in numeric results can also arise from decisions in the middle of calculations. For example, if you round off your value for the mean before you calculate the sum of squared deviations, your standard deviation probably won't agree with a computer program that calculates using many decimal places. (We do recommend that you do calculations using as many digits as you can to minimize this effect.)

Don't be overly concerned with these discrepancies, especially if the differences are small. They don't mean that your answer is "wrong," and they usually won't change any conclusion you might draw about the data. Sometimes (in footnotes and in the answers in the back of the book) we'll note alternative results, but we could never list all the possible values, so we'll rely on your common sense to focus on the meaning rather than on the digits. Remember: Answers are sentences!

TI Tips

Calculating the statistics

```

EDIT 0:00 TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
    
```

```

1-Var Stats L1
    
```

```

1-Var Stats
x=22
Σx=550
Σx²=12480
Sx=3.979112129
σx=3.898717738
n=25
    
```

```

1-Var Stats
fn=25
minX=12
Q1=19.5
Med=22
Q3=25
maxX=29
    
```

Your calculator can easily find all the numerical summaries of data. To try it out, you simply need a set of values in one of your datalists. We'll illustrate using the boys' agility test results from this chapter's earlier TI Tips (still in L1), but you can use any data currently stored in your calculator.

- Under the **STAT CALC** menu, select **1-Var Stats** and hit **ENTER**.
- Specify the location of your data, creating a command like **1-Var Stats L1**.
- Hit **ENTER** again.

Voilà! Everything you wanted to know, and more. Among all of the information shown, you are primarily interested in these statistics: \bar{x} (the mean), Sx (the standard deviation), n (the count), and—scrolling down—**minX** (the smallest datum), **Q₁** (the first quartile), **Med** (the median), **Q₃** (the third quartile), and **maxX** (the largest datum).

Sorry, but the TI doesn't explicitly tell you the range or the IQR. Just subtract: $IQR = Q_3 - Q_1 = 25 - 19.5 = 5.5$. What's the range?

By the way, if the data come as a frequency table with the values stored in, say, **L4** and the corresponding frequencies in **L5**, all you have to do is ask for **1-Var Stats L4,L5**.

WHAT CAN GO WRONG?

A data display should tell a story about the data. To do that, it must speak in a clear language, making plain what variable is displayed, what any axis shows, and what the values of the data are. And it must be consistent in those decisions.

A display of quantitative data can go wrong in many ways. The most common failures arise from only a few basic errors:

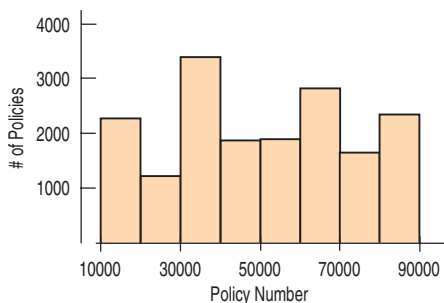


FIGURE 4.14
It's not appropriate to display these data with a histogram.

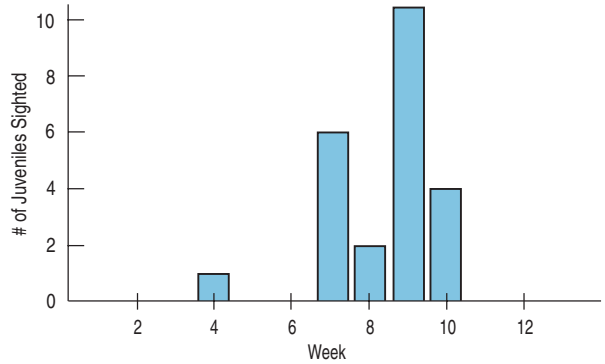
► **Don't make a histogram of a categorical variable.** Just because the variable contains numbers doesn't mean that it's quantitative. Here's a histogram of the insurance policy numbers of some workers. It's not very informative because the policy numbers are just labels. A histogram or stem-and-leaf display of a categorical variable makes no sense. A bar chart or pie chart would be more appropriate.

► **Don't look for shape, center, and spread of a bar chart.** A bar chart showing the sizes of the piles displays the distribution of a categorical variable, but the bars could be arranged in any order left to right. Concepts like symmetry, center, and spread make sense only for quantitative variables.

(continued)

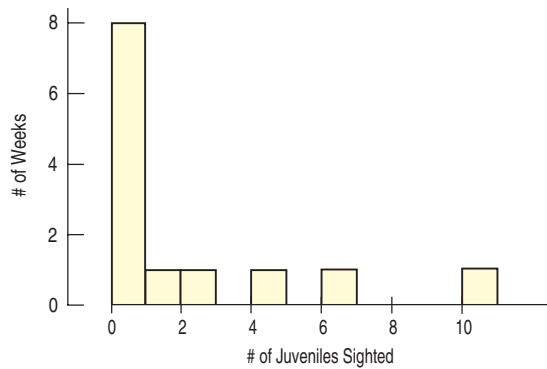
► **Don't use bars in every display—save them for histograms and bar charts.** In a bar chart, the bars indicate how many cases of a categorical variable are piled in each category. Bars in a histogram indicate the number of cases piled in each interval of a quantitative variable. In both bar charts and histograms, the bars represent counts of data values. Some people create other displays that use bars to represent individual data values. Beware: Such graphs are neither bar charts nor histograms. For example, a student was asked to make a histogram from data showing the number of juvenile bald eagles seen during each of the 13 weeks in the winter of 2003–2004 at a site in Rock Island, IL. Instead, he made this plot:

FIGURE 4.15
This isn't a histogram or a bar chart. It's an ill-conceived graph that uses bars to represent individual data values (number of eagles sighted) week by week.



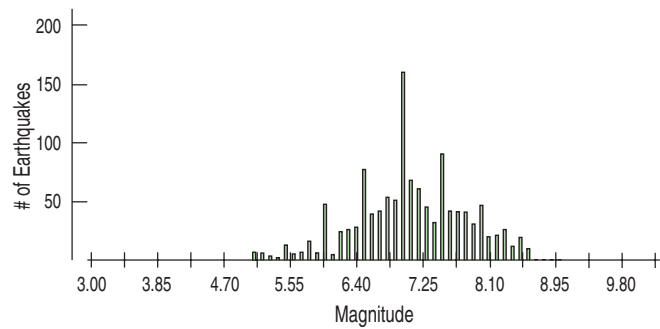
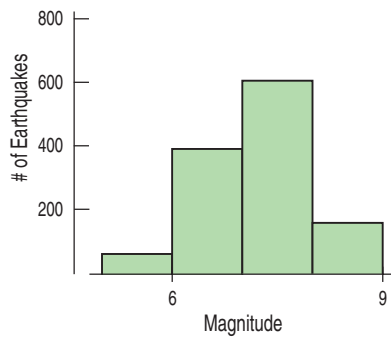
Look carefully. That's not a histogram. A histogram shows *What* we've measured along the horizontal axis and counts of the associated *Who's* represented as bar heights. This student has it backwards: He used bars to show counts of birds for each week.¹⁰ We need counts of weeks. A correct histogram should have a tall bar at "0" to show there were many weeks when no eagles were seen, like this:

FIGURE 4.16
A histogram of the eagle-sighting data shows the number of weeks in which different counts of eagles occurred. This display shows the distribution of juvenile-eagle sightings.



► **Choose a bin width appropriate to the data.** Computer programs usually do a pretty good job of choosing histogram bin widths. Often there's an easy way to adjust the width, sometimes interactively. Here are the tsunami earthquakes with two (rather extreme) choices for the bin size:

¹⁰ Edward Tufte, in his book *The Visual Display of Quantitative Information*, proposes that graphs should have a high data-to-ink ratio. That is, we shouldn't waste a lot of ink to display a single number when a dot would do the job.



The task of summarizing a quantitative variable is relatively simple, and there is a simple path to follow. However, you need to watch out for certain features of the data that make summarizing them with a number dangerous. Here’s some advice:

- ▶ **Don’t forget to do a reality check.** Don’t let the computer or calculator do your thinking for you. Make sure the calculated summaries make sense. For example, does the mean look like it is in the center of the histogram? Think about the spread: An IQR of 50 mpg would clearly be wrong for gas mileage. And no measure of spread can be negative. The standard deviation can take the value 0, but only in the very unusual case that all the data values equal the same number. If you see an IQR or standard deviation equal to 0, it’s probably a sign that something’s wrong with the data.
- ▶ **Don’t forget to sort the values before finding the median or percentiles.** It seems obvious, but when you work by hand, it’s easy to forget to sort the data first before counting in to find medians, quartiles, or other percentiles. Don’t report that the median of the five values 194, 5, 1, 17, and 893 is 1 just because 1 is the middle number.
- ▶ **Don’t worry about small differences when using different methods.** Finding the 10th percentile or the lower quartile in a data set sounds easy enough. But it turns out that the definitions are not exactly clear. If you compare different statistics packages or calculators, you may find that they give slightly different answers for the same data. These differences, though, are unlikely to be important in interpreting the data, the quartiles, or the IQR, so don’t let them worry you.

Gold Card Customers—Regions National Banks

Month	April 2007	May 2007
Average Zip Code	45,034.34	38,743.34

- ▶ **Don’t compute numerical summaries of a categorical variable.** Neither the mean zip code nor the standard deviation of social security numbers is meaningful. If the variable is categorical, you should instead report summaries such as percentages of individuals in each category. It is easy to make this mistake when using technology to do the summaries for you. After all, the computer doesn’t care what the numbers mean.


▶ **Don’t report too many decimal places.** Statistical programs and calculators often report a ridiculous number of digits. A general rule for numerical summaries is to report one or two more digits than the number of digits in the data. For example, earlier we saw a dotplot of the Kentucky Derby race times. The mean and standard deviation of those times could be reported as:

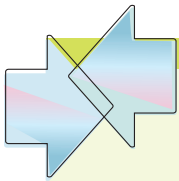
$$\bar{y} = 130.63401639344262\text{sec} \quad s = 13.66448201942662\text{sec}$$

But we knew the race times only to the nearest quarter second, so the extra digits are meaningless.

- ▶ **Don’t round in the middle of a calculation.** Don’t report too many decimal places, but it’s best not to do any rounding until the end of your calculations. Even though you might report the mean of the earthquakes as 7.08, it’s really 7.08339. Use the more precise number in your calculations if you’re finding the standard deviation by hand—or be prepared to see small differences in your final result.

(continued)

- ▶ **Watch out for multiple modes.** The summaries of the Kentucky Derby times are meaningless for another reason. As we saw in the dotplot, the Derby was initially a longer race. It would make much more sense to report that the old 1.5 mile Derby had a mean time of 159.6 seconds, while the current Derby has a mean time of 124.6 seconds. If the distribution has multiple modes, consider separating the data into different groups and summarizing each group separately.
- ▶ **Beware of outliers.** The median and IQR are resistant to outliers, but the mean and standard deviation are not. To help spot outliers . . .
- ▶ **Don't forget to: Make a picture (make a picture, make a picture).** The sensitivity of the mean and standard deviation to outliers is one reason you should always make a picture of the data. Summarizing a variable with its mean and standard deviation when you have not looked at a histogram or dotplot to check for outliers or skewness invites disaster. You may find yourself drawing absurd or dangerously wrong conclusions about the data. And, of course, you should demand no less of others. Don't accept a mean and standard deviation blindly without some evidence that the variable they summarize is unimodal, symmetric, and free of outliers. 



CONNECTIONS

Distributions of quantitative variables, like those of categorical variables, show the possible values and their relative frequencies. A histogram shows the distribution of values in a quantitative variable with adjacent bars. Don't confuse histograms with bar charts, which display categorical variables. For categorical data, the mode is the category with the biggest count. For quantitative data, modes are peaks in the histogram.

The shape of the distribution of a quantitative variable is an important concept in most of the subsequent chapters. We will be especially interested in distributions that are unimodal and symmetric.

In addition to their shape, we summarize distributions with center and spread, usually pairing a measure of center with a measure of spread: median with IQR and mean with standard deviation. We favor the mean and standard deviation when the shape is unimodal and symmetric, but choose the median and IQR for skewed distributions or when there are outliers we can't otherwise set aside.

WHAT HAVE WE LEARNED?



We've learned how to make a picture of quantitative data to help us see the story the data have to *Tell*.

- ▶ We can display the distribution of quantitative data with a *histogram*, a *stem-and-leaf* display, or a *dotplot*.
- ▶ We *Tell* what we see about the distribution by talking about *shape*, *center*, *spread*, and any *unusual features*.

We've learned how to summarize distributions of quantitative variables numerically.

- ▶ Measures of center for a distribution include the median and the mean.

We write the formula for the mean as $\bar{y} = \frac{\sum y}{n}$.

- ▶ Measures of spread include the range, IQR, and standard deviation.

The standard deviation is computed as $s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$.

The median and IQR are not usually given as formulas.

- ▶ We'll report the median and IQR when the distribution is skewed. If it's symmetric, we'll summarize the distribution with the mean and standard deviation (and possibly the median and IQR as well). Always pair the median with the IQR and the mean with the standard deviation.

We've learned to *Think* about the type of variable we're summarizing.

- ▶ All the methods of this chapter assume that the data are quantitative.
- ▶ The **Quantitative Data Condition** serves as a check that the data are, in fact, quantitative. One good way to be sure is to know the measurement units. You'll want those as part of the *Think* step of your answers.

Terms

Distribution	44. The distribution of a quantitative variable slices up all the possible values of the variable into equal-width bins and gives the number of values (or counts) falling into each bin.
Histogram (relative frequency histogram)	45. A histogram uses adjacent bars to show the distribution of a quantitative variable. Each bar represents the frequency (or relative frequency) of values falling in each bin.
Gap	45. A region of the distribution where there are no values.
Stem-and-leaf display	47. A stem-and-leaf display shows quantitative data values in a way that sketches the distribution of the data. It's best described in detail by example.
Dotplot	49. A dotplot graphs a dot for each case against a single axis.
Shape	49. To describe the shape of a distribution, look for <ul style="list-style-type: none"> ▶ single vs. multiple modes. ▶ symmetry vs. skewness. ▶ outliers and gaps.
Center	52, 58. The place in the distribution of a variable that you'd point to if you wanted to attempt the impossible by summarizing the entire distribution with a single number. Measures of center include the mean and median.
Spread	54, 61. A numerical summary of how tightly the values are clustered around the center. Measures of spread include the IQR and standard deviation.
Mode	49. A hump or local high point in the shape of the distribution of a variable. The apparent location of modes can change as the scale of a histogram is changed.
Unimodal (Bimodal)	50. Having one mode. This is a useful term for describing the shape of a histogram when it's generally mound-shaped. Distributions with two modes are called bimodal . Those with more than two are multimodal .
Uniform	50. A distribution that's roughly flat is said to be uniform.
Symmetric	50. A distribution is symmetric if the two halves on either side of the center look approximately like mirror images of each other.
Tails	50. The tails of a distribution are the parts that typically trail off on either side. Distributions can be characterized as having long tails (if they straggle off for some distance) or short tails (if they don't).
Skewed	50. A distribution is skewed if it's not symmetric and one tail stretches out farther than the other. Distributions are said to be skewed left when the longer tail stretches to the left, and skewed right when it goes to the right.
Outliers	51. Outliers are extreme values that don't appear to belong with the rest of the data. They may be unusual values that deserve further investigation, or they may be just mistakes; there's no obvious way to tell. Don't delete outliers automatically—you have to think about them. Outliers can affect many statistical analyses, so you should always be alert for them.
Median	52. The median is the middle value, with half of the data above and half below it. If n is even, it is the average of the two middle values. It is usually paired with the IQR.
Range	54. The difference between the lowest and highest values in a data set. $Range = max - min$.
Quartile	54. The lower quartile (Q1) is the value with a quarter of the data below it. The upper quartile (Q3) has three quarters of the data below it. The median and quartiles divide data into four parts with equal numbers of data values.

Interquartile range (IQR)	55. The IQR is the difference between the first and third quartiles. $IQR = Q3 - Q1$. It is usually reported along with the median.
Percentile	55. The i th percentile is the number that falls above $i\%$ of the data.
5-Number Summary	56. The 5-number summary of a distribution reports the minimum value, $Q1$, the median, $Q3$, and the maximum value.
Mean	58. The mean is found by summing all the data values and dividing by the count: $\bar{y} = \frac{Total}{n} = \frac{\sum y}{n}.$ It is usually paired with the standard deviation.
Resistant	59. A calculated summary is said to be resistant if outliers have only a small effect on it.
Variance	61. The variance is the sum of squared deviations from the mean, divided by the count minus 1: $s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}.$ It is useful in calculations later in the book.
Standard deviation	61. The standard deviation is the square root of the variance: $s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$ It is usually reported along with the mean.

Skills

THINK

- ▶ Be able to identify an appropriate display for any quantitative variable.
- ▶ Be able to guess the shape of the distribution of a variable by knowing something about the data.
- ▶ Be able to select a suitable measure of center and a suitable measure of spread for a variable based on information about its distribution.
- ▶ Know the basic properties of the median: The median divides the data into the half of the data values that are below the median and the half that are above.
- ▶ Know the basic properties of the mean: The mean is the point at which the histogram balances.
- ▶ Know that the standard deviation summarizes how spread out all the data are around the mean.
- ▶ Understand that the median and IQR resist the effects of outliers, while the mean and standard deviation do not.
- ▶ Understand that in a skewed distribution, the mean is pulled in the direction of the skewness (toward the longer tail) relative to the median.

SHOW

- ▶ Know how to display the distribution of a quantitative variable with a stem-and-leaf display (drawn by hand for smaller data sets), a dotplot, or a histogram (made by computer for larger data sets).
- ▶ Know how to compute the mean and median of a set of data.
- ▶ Know how to compute the standard deviation and IQR of a set of data.

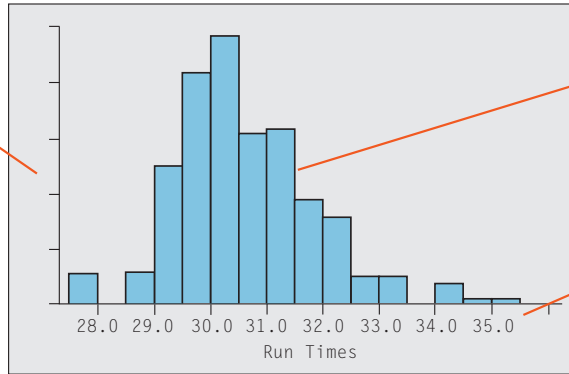
TELL

- ▶ Be able to describe the distribution of a quantitative variable in terms of its shape, center, and spread.
- ▶ Be able to describe any anomalies or extraordinary features revealed by the display of a variable.
- ▶ Know how to describe summary measures in a sentence. In particular, know that the common measures of center and spread have the same units as the variable that they summarize, and should be described in those units.
- ▶ Be able to describe the distribution of a quantitative variable with a description of the shape of the distribution, a numerical measure of center, and a numerical measure of spread. Be sure to note any unusual features, such as outliers, too.

DISPLAYING AND SUMMARIZING QUANTITATIVE VARIABLES ON THE COMPUTER

Almost any program that displays data can make a histogram, but some will do a better job of determining where the bars should start and how they should partition the span of the data.

The vertical scale may be counts or proportions. Sometimes it isn't clear which. But the shape of the histogram is the same either way.



Most packages choose the number of bars for you automatically. Often you can adjust that choice.

The axis should be clearly labeled so you can tell what "pile" each bar represents. You should be able to tell the lower and upper bounds of each bar.

Many statistics packages offer a prepackaged collection of summary measures. The result might look like this:

Variable: W eight
 N = 234
 Mean = 143.3 Median = 139
 St. Dev = 11.1 IQR = 14

Alternatively, a package might make a table for several variables and summary measures:

AS **Case Study: Describing Distribution Shapes.** Who's safer in a crash—passengers or the driver? Investigate with your statistics package.

Variable	N	mean	median	stdev	IQR
Weight	234	143.3	139	11.1	14
Height	234	68.3	68.1	4.3	5
Score	234	86	88	9	5

It is usually easy to read the results and identify each computed summary. You should be able to read the summary statistics produced by any computer package.

Packages often provide many more summary statistics than you need. Of course, some of these may not be appropriate when the data are skewed or have outliers. It is your responsibility to check a histogram or stem-and-leaf display and decide which summary statistics to use.

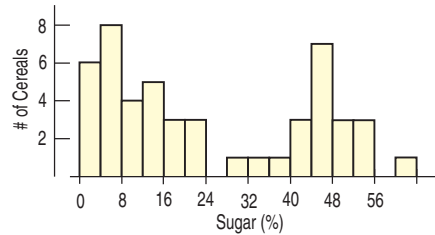
It is common for packages to report summary statistics to many decimal places of "accuracy." Of course, it is rare data that have such accuracy in the original measurements. The ability to calculate to six or seven digits beyond the decimal point doesn't mean that those digits have any meaning. Generally it's a good idea to round these values, allowing perhaps one more digit of precision than was given in the original data.

Displays and summaries of quantitative variables are among the simplest things you can do in most statistics packages.

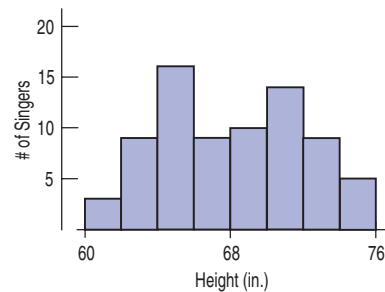
EXERCISES

- Histogram.** Find a histogram that shows the distribution of a variable in a newspaper, a magazine, or the Internet.
 - Does the article identify the W's?
 - Discuss whether the display is appropriate.
 - Discuss what the display reveals about the variable and its distribution.
 - Does the article accurately describe and interpret the data? Explain.
- Not a histogram.** Find a graph other than a histogram that shows the distribution of a quantitative variable in a newspaper, a magazine, or the Internet.
 - Does the article identify the W's?
 - Discuss whether the display is appropriate for the data.
 - Discuss what the display reveals about the variable and its distribution.
 - Does the article accurately describe and interpret the data? Explain.
- In the news.** Find an article in a newspaper, a magazine, or the Internet that discusses an "average."
 - Does the article discuss the W's for the data?
 - What are the units of the variable?
 - Is the average used the median or the mean? How can you tell?
 - Is the choice of median or mean appropriate for the situation? Explain.
- In the news II.** Find an article in a newspaper, a magazine, or the Internet that discusses a measure of spread.
 - Does the article discuss the W's for the data?
 - What are the units of the variable?
 - Does the article use the range, IQR, or standard deviation?
 - Is the choice of measure of spread appropriate for the situation? Explain.
- Thinking about shape.** Would you expect distributions of these variables to be uniform, unimodal, or bimodal? Symmetric or skewed? Explain why.
 - The number of speeding tickets each student in the senior class of a college has ever had.
 - Players' scores (number of strokes) at the U.S. Open golf tournament in a given year.
 - Weights of female babies born in a particular hospital over the course of a year.
 - The length of the average hair on the heads of students in a large class.
- More shapes.** Would you expect distributions of these variables to be uniform, unimodal, or bimodal? Symmetric or skewed? Explain why.
 - Ages of people at a Little League game.
 - Number of siblings of people in your class.
 - Pulse rates of college-age males.
 - Number of times each face of a die shows in 100 tosses.

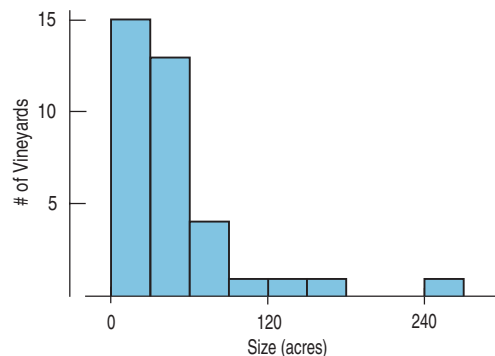
- T 7. Sugar in cereals.** The histogram displays the sugar content (as a percent of weight) of 49 brands of breakfast cereals.



- Describe this distribution.
 - What do you think might account for this shape?
- T 8. Singers.** The display shows the heights of some of the singers in a chorus, collected so that the singers could be positioned on stage with shorter ones in front and taller ones in back.

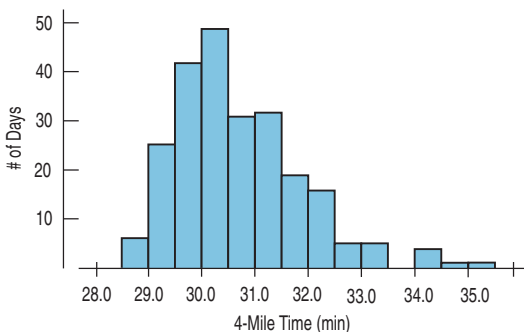


- Describe the distribution.
 - Can you account for the features you see here?
- T 9. Vineyards.** The histogram shows the sizes (in acres) of 36 vineyards in the Finger Lakes region of New York.



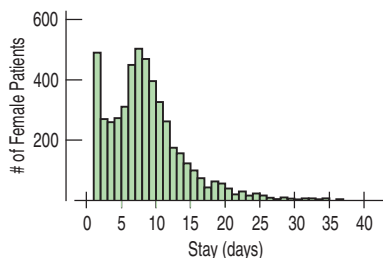
- Approximately what percentage of these vineyards are under 60 acres?
- Write a brief description of this distribution (shape, center, spread, unusual features).

10. **Run times.** One of the authors collected the times (in minutes) it took him to run 4 miles on various courses during a 10-year period. Here is a histogram of the times.



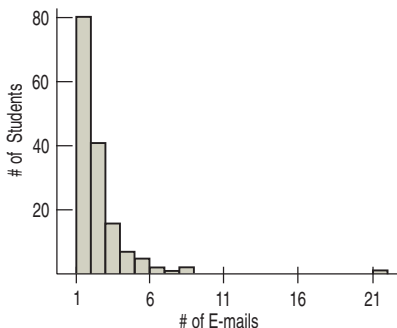
Describe the distribution and summarize the important features. What is it about running that might account for the shape you see?

11. **Heart attack stays.** The histogram shows the lengths of hospital stays (in days) for all the female patients admitted to hospitals in New York during one year with a primary diagnosis of acute myocardial infarction (heart attack).



- From the histogram, would you expect the mean or median to be larger? Explain.
- Write a few sentences describing this distribution (shape, center, spread, unusual features).
- Which summary statistics would you choose to summarize the center and spread in these data? Why?

- T** 12. **E-mails.** A university teacher saved every e-mail received from students in a large Introductory Statistics class during an entire term. He then counted, for each student who had sent him at least one e-mail, how many e-mails each student had sent.



- From the histogram, would you expect the mean or the median to be larger? Explain.
- Write a few sentences describing this distribution (shape, center, spread, unusual features).

- Which summary statistics would you choose to summarize the center and spread in these data? Why?

13. **Super Bowl points.** How many points do football teams score in the Super Bowl? Here are the total numbers of points scored by both teams in each of the first 42 Super Bowl games:

45, 47, 23, 30, 29, 27, 21, 31, 22, 38, 46, 37, 66, 50, 37, 47, 44, 47, 54, 56, 59, 52, 36, 65, 39, 61, 69, 43, 75, 44, 56, 55, 53, 39, 41, 37, 69, 61, 45, 31, 46, 31

- Find the median.
 - Find the quartiles.
 - Write a description based on the 5-number summary.
14. **Super Bowl wins.** In the Super Bowl, by how many points does the winning team outscore the losers? Here are the winning margins for the first 42 Super Bowl games:

25, 19, 9, 16, 3, 21, 7, 17, 10, 4, 18, 17, 4, 12, 17, 5, 10, 29, 22, 36, 19, 32, 4, 45, 1, 13, 35, 17, 23, 10, 14, 7, 15, 7, 27, 3, 27, 3, 3, 11, 12, 3

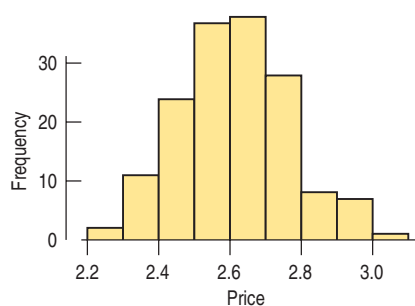
- Find the median.
 - Find the quartiles.
 - Write a description based on the 5-number summary.
15. **Standard deviation I.** For each lettered part, a through c, examine the two given sets of numbers. Without doing any calculations, decide which set has the larger standard deviation and explain why. Then check by finding the standard deviations *by hand*.

Set 1	Set 2
a) 3, 5, 6, 7, 9	2, 4, 6, 8, 10
b) 10, 14, 15, 16, 20	10, 11, 15, 19, 20
c) 2, 6, 6, 9, 11, 14	82, 86, 86, 89, 91, 94

16. **Standard deviation II.** For each lettered part, a through c, examine the two given sets of numbers. Without doing any calculations, decide which set has the larger standard deviation and explain why. Then check by finding the standard deviations *by hand*.

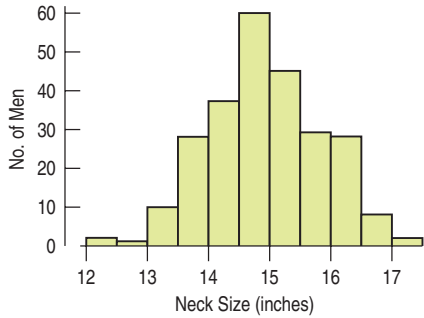
Set 1	Set 2
a) 4, 7, 7, 7, 10	4, 6, 7, 8, 10
b) 100, 140, 150, 160, 200	10, 50, 60, 70, 110
c) 10, 16, 18, 20, 22, 28	48, 56, 58, 60, 62, 70

- T** 17. **Pizza prices.** The histogram shows the distribution of the prices of plain pizza slices (in \$) for 156 weeks in Dallas, TX.



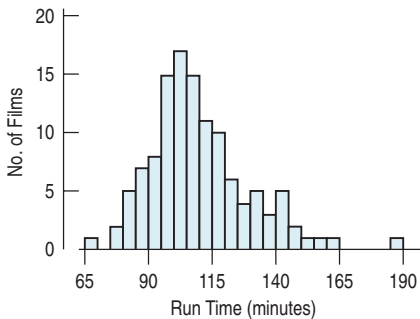
Which summary statistics would you choose to summarize the center and spread in these data? Why?

- T 18. Neck size.** The histogram shows the neck sizes (in inches) of 250 men recruited for a health study in Utah.



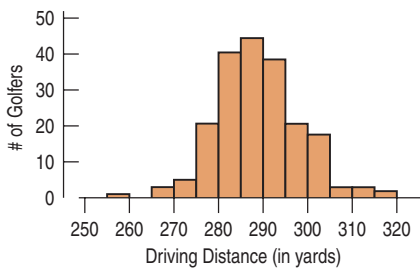
Which summary statistics would you choose to summarize the center and spread in these data? Why?

- T 19. Pizza prices again.** Look again at the histogram of the pizza prices in Exercise 17.
- Is the mean closer to \$2.40, \$2.60, or \$2.80? Why?
 - Is the standard deviation closer to \$0.15, \$0.50, or \$1.00? Explain.
- T 20. Neck sizes again.** Look again at the histogram of men's neck sizes in Exercise 18.
- Is the mean closer to 14, 15, or 16 inches? Why?
 - Is the standard deviation closer to 1 inch, 3 inches, or 5 inches? Explain.
- T 21. Movie lengths.** The histogram shows the running times in minutes of 122 feature films released in 2005.



- You plan to see a movie this weekend. Based on these movies, how long do you expect a typical movie to run?
- Would you be surprised to find that your movie ran for $2\frac{1}{2}$ hours (150 minutes)?
- Which would you expect to be higher: the mean or the median run time for all movies? Why?

- T 22. Golf drives.** The display shows the average drive distance (in yards) for 202 professional golfers on the men's PGA tour.



- Describe this distribution.
- Approximately what proportion of professional male golfers drive, on average, less than 280 yards?
- Estimate the mean by examining the histogram.
- Do you expect the mean to be smaller than, approximately equal to, or larger than the median? Why?

- 23. Movie lengths II.** Exercise 21 looked at the running times of movies released in 2005. The standard deviation of these running times is 19.6 minutes, and the quartiles are $Q_1 = 97$ minutes and $Q_3 = 119$ minutes.

- Write a sentence or two describing the spread in running times based on
 - the quartiles.
 - the standard deviation.
- Do you have any concerns about using either of these descriptions of spread? Explain.

- 24. Golf drives II.** Exercise 22 looked at distances PGA golfers can hit the ball. The standard deviation of these average drive distances is 9.3 yards, and the quartiles are $Q_1 = 282$ yards and $Q_3 = 294$ yards.

- Write a sentence or two describing the spread in distances based on
 - the quartiles.
 - the standard deviation.
- Do you have any concerns about using either of these descriptions of spread? Explain.

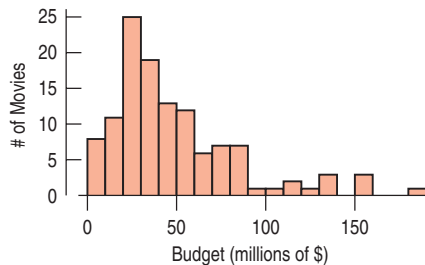
- 25. Mistake.** A clerk entering salary data into a company spreadsheet accidentally put an extra "0" in the boss's salary, listing it as \$2,000,000 instead of \$200,000. Explain how this error will affect these summary statistics for the company payroll:

- measures of center: median and mean.
- measures of spread: range, IQR, and standard deviation.

- 26. Cold weather.** A meteorologist preparing a talk about global warming compiled a list of weekly low temperatures (in degrees Fahrenheit) he observed at his southern Florida home last year. The coldest temperature for any week was 36°F , but he inadvertently recorded the Celsius value of 2° . Assuming that he correctly listed all the other temperatures, explain how this error will affect these summary statistics:

- measures of center: mean and median.
- measures of spread: range, IQR, and standard deviation.

- T 27. Movie budgets.** The histogram shows the budgets (in millions of dollars) of major release movies in 2005.



An industry publication reports that the average movie costs \$35 million to make, but a watchdog group con-

cerned with rising ticket prices says that the average cost is \$46.8 million. What statistic do you think each group is using? Explain.

28. **Sick days.** During contract negotiations, a company seeks to change the number of sick days employees may take, saying that the annual “average” is 7 days of absence per employee. The union negotiators counter that the “average” employee misses only 3 days of work each year. Explain how both sides might be correct, identifying the measure of center you think each side is using and why the difference might exist.
29. **Payroll.** A small warehouse employs a supervisor at \$1200 a week, an inventory manager at \$700 a week, six stock boys at \$400 a week, and four drivers at \$500 a week.
- Find the mean and median wage.
 - How many employees earn more than the mean wage?
 - Which measure of center best describes a typical wage at this company: the mean or the median?
 - Which measure of spread would best describe the payroll: the range, the IQR, or the standard deviation? Why?
30. **Singers.** The frequency table shows the heights (in inches) of 130 members of a choir.

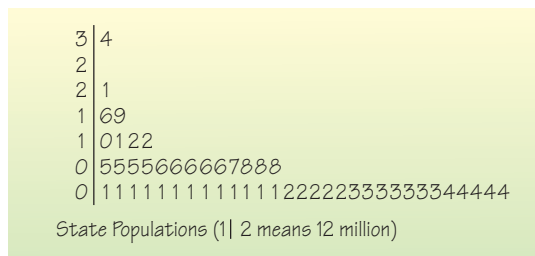
Height	Count	Height	Count
60	2	69	5
61	6	70	11
62	9	71	8
63	7	72	9
64	5	73	4
65	20	74	2
66	18	75	4
67	7	76	1
68	12		

- Find the median and IQR.
 - Find the mean and standard deviation.
 - Display these data with a histogram.
 - Write a few sentences describing the distribution.
31. **Gasoline.** In March 2006, 16 gas stations in Grand Junction, CO, posted these prices for a gallon of regular gasoline:

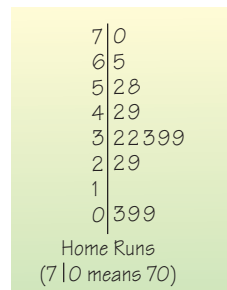
2.22	2.21	2.45	2.24
2.27	2.28	2.27	2.23
2.26	2.46	2.29	2.32
2.36	2.38	2.33	2.27

- Make a stem-and-leaf display of these gas prices. Use split stems; for example, use two 2.2 stems—one for prices between \$2.20 and \$2.24 and the other for prices from \$2.25 to \$2.29.
- Describe the shape, center, and spread of this distribution.
- What unusual feature do you see?

32. **The Great One.** During his 20 seasons in the NHL, Wayne Gretzky scored 50% more points than anyone who ever played professional hockey. He accomplished this amazing feat while playing in 280 fewer games than Gordie Howe, the previous record holder. Here are the number of games Gretzky played during each season:
- 79, 80, 80, 80, 74, 80, 80, 79, 64, 78, 73, 78, 74, 45, 81, 48, 80, 82, 82, 70
- Create a stem-and-leaf display for these data, using split stems.
 - Describe the shape of the distribution.
 - Describe the center and spread of this distribution.
 - What unusual feature do you see? What might explain this?
33. **States.** The stem-and-leaf display shows populations of the 50 states and Washington, DC, in millions of people, according to the 2000 census.



- What measures of center and spread are most appropriate?
 - Without doing any calculations, which must be larger: the median or the mean? Explain how you know.
 - From the stem-and-leaf display, find the median and the interquartile range.
 - Write a few sentences describing this distribution.
34. **Wayne Gretzky.** In Exercise 32, you examined the number of games played by hockey great Wayne Gretzky during his 20-year career in the NHL.
- Would you use the median or the mean to describe the center of this distribution? Why?
 - Find the median.
 - Without actually finding the mean, would you expect it to be higher or lower than the median? Explain.
35. **Home runs.** The stem-and-leaf display shows the number of home runs hit by Mark McGwire during the 1986–2001 seasons. Describe the distribution, mentioning its shape and any unusual features.



36. **Bird species.** The Cornell Lab of Ornithology holds an annual Christmas Bird Count (www.birdsource.org), in which bird watchers at various locations around the country see how many different species of birds they can spot. Here are some of the counts reported from sites in Texas during the 1999 event:

228	178	186	162	206	166	163
183	181	206	177	175	167	162
160	160	157	156	153	153	152

- Create a stem-and-leaf display of these data.
- Write a brief description of the distribution. Be sure to discuss the overall shape as well as any unusual features.

37. **Hurricanes 2006.** The data below give the number of hurricanes classified as major hurricanes in the Atlantic Ocean each year from 1944 through 2006, as reported by NOAA (www.nhc.noaa.gov):

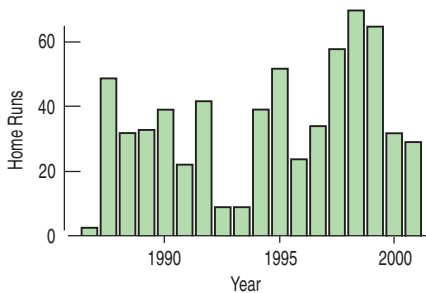
3, 2, 1, 2, 4, 3, 7, 2, 3, 3, 2, 5, 2, 4, 2, 2, 6, 0, 2, 5, 1, 3, 1, 0, 3, 2, 1, 0, 1, 2, 3, 2, 1, 2, 2, 2, 3, 1, 1, 1, 3, 0, 1, 3, 2, 1, 2, 1, 1, 0, 5, 6, 1, 3, 5, 3, 3, 2, 3, 6, 7, 2

- Create a dotplot of these data.
- Describe the distribution.

38. **Horsepower.** Create a stem-and-leaf display for these horsepowers of autos reviewed by *Consumer Reports* one year, and describe the distribution:

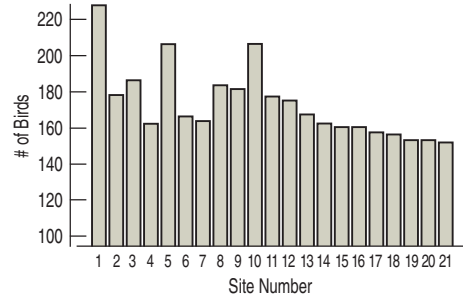
155	103	130	80	65
142	125	129	71	69
125	115	138	68	78
150	133	135	90	97
68	105	88	115	110
95	85	109	115	71
97	110	65	90	
75	120	80	70	

39. **Home runs again.** Students were asked to make a histogram of the number of home runs hit by Mark McGwire from 1986 to 2001 (see Exercise 35). One student submitted the following display:



- Comment on this graph.
- Create your own histogram of the data.

40. **Return of the birds.** Students were given the assignment to make a histogram of the data on bird counts reported in Exercise 36. One student submitted the following display:



- Comment on this graph.
- Create your own histogram of the data.

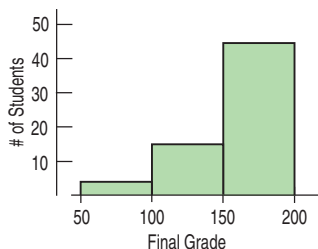
41. **Acid rain.** Two researchers measured the pH (a scale on which a value of 7 is neutral and values below 7 are acidic) of water collected from rain and snow over a 6-month period in Allegheny County, PA. Describe their data with a graph and a few sentences:

4.57	5.62	4.12	5.29	4.64	4.31	4.30	4.39	4.45
5.67	4.39	4.52	4.26	4.26	4.40	5.78	4.73	4.56
5.08	4.41	4.12	5.51	4.82	4.63	4.29	4.60	

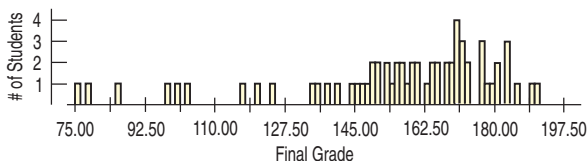
42. **Marijuana 2003.** In 2003 the Council of Europe published a report entitled *The European School Survey Project on Alcohol and Other Drugs* (www.espad.org). Among other issues, the survey investigated the percentages of 16-year-olds who had used marijuana. Shown here are the results for 20 European countries. Create an appropriate graph of these data, and describe the distribution.

Country	Percentage	Country	Percentage
Austria	21%	Italy	27%
Belgium	32%	Latvia	16%
Bulgaria	21%	Lithuania	13%
Croatia	22%	Malta	10%
Cyprus	4%	Netherlands	28%
Czech Republic	44%	Norway	9%
Denmark	23%	Poland	18%
Estonia	23%	Portugal	15%
Faroe Islands	9%	Romania	3%
Finland	11%	Russia	22%
France	22%	Slovak Republic	27%
Germany	27%	Slovenia	28%
Greece	6%	Sweden	7%
Greenland	27%	Switzerland	40%
Hungary	16%	Turkey	4%
Iceland	13%	Ukraine	21%
Ireland	39%	United Kingdom	38%
Isle of Man	39%		

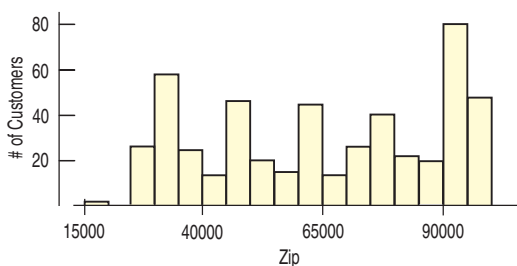
43. **Final grades.** A professor (of something other than Statistics!) distributed the following histogram to show the distribution of grades on his 200-point final exam. Comment on the display.



44. **Final grades revisited.** After receiving many complaints about his final-grade histogram from students currently taking a Statistics course, the professor from Exercise 43 distributed the following revised histogram:



- a) Comment on this display.
 b) Describe the distribution of grades.
45. **Zip codes.** Holes-R-U's, an Internet company that sells piercing jewelry, keeps transaction records on its sales. At a recent sales meeting, one of the staff presented a histogram of the zip codes of the last 500 customers, so that the staff might understand where sales are coming from. Comment on the usefulness and appropriateness of the display.



46. **Zip codes revisited.** Here are some summary statistics to go with the histogram of the zip codes of 500 customers from the Holes-R-U's Internet Jewelry Salon that we saw in Exercise 45:

Count	500
Mean	64,970.0
StdDev	23,523.0
Median	64,871
IQR	44,183
Q1	46,050
Q3	90,233

What can these statistics tell you about the company's sales?

47. **Math scores 2005.** The National Center for Education Statistics (<http://nces.ed.gov/nationsreportcard/>) reported 2005 average mathematics achievement scores for eighth graders in all 50 states:

State	Score	State	Score
Alabama	225	Montana	241
Alaska	236	Nebraska	238
Arizona	230	Nevada	230
Arkansas	236	New Hampshire	246
California	230	New Jersey	244
Colorado	239	New Mexico	224
Connecticut	242	New York	238
Delaware	240	North Carolina	241
Florida	239	North Dakota	243
Georgia	234	Ohio	242
Hawaii	230	Oklahoma	234
Idaho	242	Oregon	238
Illinois	233	Pennsylvania	241
Indiana	240	Rhode Island	233
Iowa	240	South Carolina	238
Kansas	246	South Dakota	242
Kentucky	231	Tennessee	232
Louisiana	230	Texas	242
Maine	241	Utah	239
Maryland	238	Vermont	244
Massachusetts	247	Virginia	240
Michigan	238	Washington	242
Minnesota	246	West Virginia	231
Mississippi	227	Wisconsin	241
Missouri	235	Wyoming	243

- a) Find the median, the IQR, the mean, and the standard deviation of these state averages.
 b) Which summary statistics would you report for these data? Why?
 c) Write a brief summary of the performance of eighth graders nationwide.

48. **Boomtowns.** In 2006, *Inc.* magazine (www.inc.com) listed its choice of "boomtowns" in the United States—larger cities that are growing rapidly. Here is the magazine's top 20, along with their job growth percentages:

City	1-Year Job Growth (%)
Las Vegas, NV	7.5
Fort Lauderdale, FL	4.2
Orlando, FL	4.5
West Palm Beach-Boca Raton, FL	3.4
San Bernadino-Riverside, CA	1.9
Phoenix, AZ	4.4
Northern Virginia, VA	3.1
Washington, DC-Arlington-Alexandria, VA	3.2
Tampa-St. Petersburg, FL	2.6
Camden-Burlington counties, NJ	2.6

(continued)

City	1-Year Job Growth (%)
Jacksonville, FL	2.6
Charlotte, NC	3.3
Raleigh-Cary, NC	2.8
Richmond, VA	2.9
Salt Lake City, UT	3.3
Putnam-Rockland-Westchester counties, New York	2.3
Santa Ana-Anaheim-Irvine, CA	1.7
Miami-Miami Beach, FL	2.2
Sacramento, CA	1.5
San Diego, CA	1.4

Massachusetts	458.5	Oklahoma	614.2
Michigan	482.0	Oregon	418.4
Minnesota	527.7	Pennsylvania	386.8
Mississippi	558.5	Rhode Island	454.6
Missouri	550.5	South Carolina	578.6
Montana	544.4	South Dakota	564.4
Nebraska	470.1	Tennessee	552.5
Nevada	367.9	Texas	532.7
New Hampshire	544.4	Utah	460.6
New Jersey	488.2	Vermont	545.5
New Mexico	508.8	Virginia	526.9
New York	293.4	Washington	423.6
North Carolina	505.0	West Virginia	426.7
North Dakota	553.7	Wisconsin	449.8
Ohio	451.1	Wyoming	615.0

- a) Make a suitable display of the growth rates.
 - b) Summarize the typical growth rate among these cities with a median and mean. Why do they differ?
 - c) Given what you know about the distribution, which of the measures in b) does the better job of summarizing the growth rates? Why?
 - d) Summarize the spread of the growth rate distribution with a standard deviation and with an IQR.
 - e) Given what you know about the distribution, which of the measures in d) does the better job of summarizing the growth rates? Why?
 - f) Suppose we subtract from each of the preceding growth rates the predicted U.S. average growth rate of 1.20%, so that we can look at how much these growth rates exceed the U.S. rate. How would this change the values of the summary statistics you calculated above? (*Hint: You need not recompute any of the summary statistics from scratch.*)
 - g) If we were to omit Las Vegas from the data, how would you expect the mean, median, standard deviation, and IQR to change? Explain your expectations for each.
 - h) Write a brief report about all of these growth rates.
- T 49. Gasoline usage 2004.** The California Energy Commission (www.energy.ca.gov/gasoline/) collects data on the amount of gasoline sold in each state. The following data show the per capita (gallons used per person) consumption in the year 2004. Using appropriate graphical displays and summary statistics, write a report on the gasoline use by state in the year 2004.

State	Gallons per Capita	State	Gallons per Capita
Alabama	529.4	Hawaii	358.7
Alaska	461.7	Idaho	454.8
Arizona	381.9	Illinois	408.3
Arkansas	512.0	Indiana	491.7
California	414.4	Iowa	555.1
Colorado	435.7	Kansas	511.8
Connecticut	435.7	Kentucky	526.6
Delaware	541.6	Louisiana	507.8
Florida	496.0	Maine	576.3
Georgia	537.1	Maryland	447.5

- T 50. Prisons 2005.** A report from the U.S. Department of Justice (www.ojp.usdoj.gov/bjs/) reported the percent changes in federal prison populations in 21 northeastern and midwestern states during 2005. Using appropriate graphical displays and summary statistics, write a report on the changes in prison populations.

State	Percent Change	State	Percent Change
Connecticut	-0.3	Iowa	2.5
Maine	0.0	Kansas	1.1
Massachusetts	5.5	Michigan	1.4
New Hampshire	3.3	Minnesota	6.0
New Jersey	2.2	Missouri	-0.8
New York	-1.6	Nebraska	7.9
Pennsylvania	3.5	North Dakota	4.4
Rhode Island	6.5	Ohio	2.3
Vermont	5.6	South Dakota	11.9
Illinois	2.0	Wisconsin	-1.0
Indiana	1.9		



JUST CHECKING **Answers**

(Thoughts will vary.)

- 1.** Roughly symmetric, slightly skewed to the right. Center around 3 miles? Few over 10 miles.
- 2.** Bimodal. Center between 1 and 2 hours? Many people watch no football; others watch most of one or more games. Probably only a few values over 5 hours.
- 3.** Strongly skewed to the right, with almost everyone at \$0; a few small prizes, with the winner an outlier.
- 4.** Fairly symmetric, somewhat uniform, perhaps slightly skewed to the right. Center in the 40s? Few ages below 25 or above 70.
- 5.** Uniform, symmetric. Center near 5. Roughly equal counts for each digit 0–9.
- 6.** Incomes are probably skewed to the right and not symmetric, making the median the more appropriate measure of center. The mean will be influenced by the high end of family incomes and not reflect the “typical” family income as well as the median would. It will give the impression that the typical income is higher than it is.
- 7.** An IQR of 30 mpg would mean that only 50% of the cars get gas mileages in an interval 30 mpg wide. Fuel economy doesn’t vary that much. 3 mpg is reasonable. It seems plausible that 50% of the cars will be within about 3 mpg of each other. An IQR of 0.3 mpg would mean that the gas mileage of half the cars varies little from the estimate. It’s unlikely that cars, drivers, and driving conditions are that consistent.
- 8.** We’d prefer a standard deviation of 2 months. Making a consistent product is important for quality. Customers want to be able to count on the MP3 player lasting somewhere close to 5 years, and a standard deviation of 2 years would mean that life-spans were highly variable.