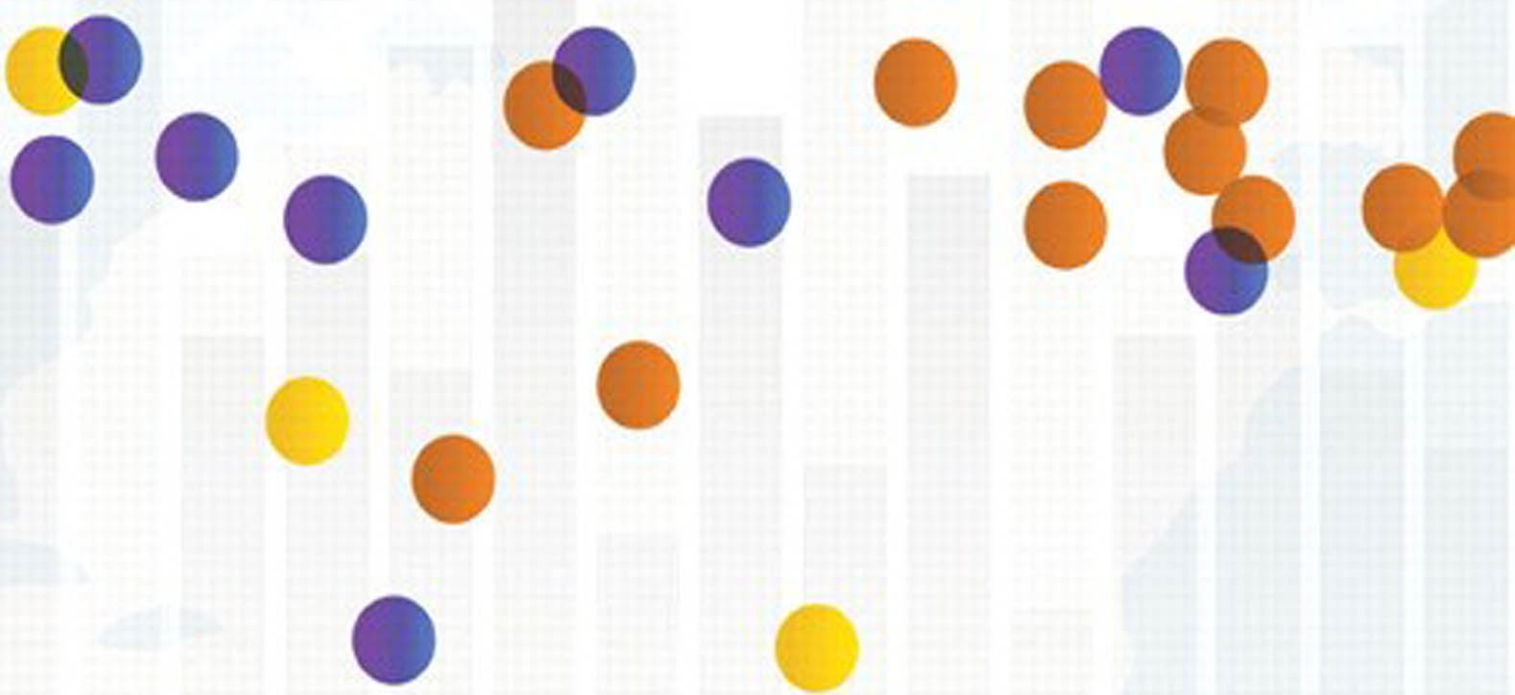
STATS

DATA & MODELS



De VEAUX VELLEMAN BOCK

4TH EDITION

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	$\hat{p}$	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$	18
		1-Proportion *z*-Test				$\sqrt{\frac{p_0 q_0}{n}}$	19
	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	$\bar{y}$	$\frac{s}{\sqrt{n}}$	20
	Two independent groups	2-Sample *t*-Test 2-Sample *t*-Interval	t df from technology	$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	22
	n Matched pairs	Paired *t*-Test Paired *t*-Interval	t df $= n - 1$	μ_d	$\bar{d}$	$\frac{s_d}{\sqrt{n}}$	23
Distributions (one categorical variable)	One sample	Goodness of fit	χ^2 df $= cells - 1$	$\sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$			24
	Many independent groups	Homogeneity χ^2 Test	χ^2 df $= (r-1)(c-1)$				
Independence (two categorical variables)	One sample	Independence χ^2 Test					
Association (two quantitative variables)	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)	
		Confidence Interval for μ_y		μ_y	$\hat{y}_y$	$\sqrt{SE^2(b_1)\cdot(x_y - \bar{x})^2 + \frac{s_e^2}{n}}$	25
		Prediction Interval for y_y		y_y	$\hat{y}_y$	$\sqrt{SE^2(b_1)\cdot(x_y - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$	
Inference about?	One group or two?	Procedure	Model	Parameter	Estimate	SE	Chapter

edition 4

Stats: Data and Models

Richard D. De Veaux
Williams College

Paul F. Velleman
Cornell University

David E. Bock
Cornell University

PEARSON

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

Editor-in-Chief: *Deirdre Lynch*
Acquisitions Editor: *Patrick Barbera*
Editorial Assistants: *Salena Casha and Justin Billing*
Program Manager: *Chere Bemelmans*
Project Manager: *Shannon Steed*
Program Management Team Lead: *Marianne Stepanian*
Project Management Team Lead: *Christina Lepre*
Media Producer: *Stephanie Green*
TestGen Content Manager: *John Flanagan*
MathXL Content Developer: *Bob Carroll*
Senior Marketing Manager: *Erin Kelly*
Senior Author Support/Technology Specialist: *Joe Vetere*
Rights and Permissions Project Manager: *Diahanne Lucas*
Procurement Specialist: *Carol Melville*
Associate Director of Design: *Andrea Nix*
Senior Designer: *Barbara Atkinson*
Text Design: *Studio Montage*
Production Management, Composition, and Illustrations: *Lumina Datamatics, Inc.*
Cover Design: *Studio Montage/Barbara Atkinson*
Cover Image: *liravega/Shutterstock*

Acknowledgements of third party content appear on pages A-51–A-52, which constitutes an extension of this copyright page.

Library of Congress Cataloging-in-Publication Data

De Veaux, Richard D.
Stats : data and models / Richard D. De Veaux, Williams College, Paul F. Velleman, Cornell University, David E. Bock, Cornell University. —4th edition.
pages cm
Includes index.
ISBN 978-0-321-98649-8
1. Statistics—Textbooks. 2. Mathematical statistics—Textbooks. I. Velleman, Paul F., 1949- II. Bock, David E. III. Title.
QA276.12.D417 2016
519.5—dc23

2014019094

1 2 3 4 5 6 7 8 9 10—RRD—18 17 16 15 14

ISBN 13: 978-0-321-99028-0 (Instructor's Edition)
ISBN 10: 0-321-99028-5 (Instructor's Edition)

ISBN 13: 978-0-321-98649-8 (Student Edition)
ISBN 10: 0-321-98649-0 (Student Edition)

PEARSON

www.pearsonhighered.com

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

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 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

Meet the Authors

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 (ASA) and an elected member of the International Statistical Institute (ISI). In 2008, he was named Statistician of the Year by the Boston Chapter of the ASA. Dick is also well known in industry, where for more than 25 years he has consulted for such Fortune 500 companies as American Express, Hewlett-Packard, Alcoa, DuPont, Pillsbury, General Electric, and Chemical Bank. Because he consulted with Mickey Hart on his book *Planet Drum*, 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 dance with Inga Weiss and Statistics 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 of the "Diminished Faculty," an a cappella Doo-Wop quartet at Williams College and sings bass in the college concert choir. Dick is the father of four children.

Paul F. Velleman has an international reputation for innovative Statistics education. He is the author and designer of the multimedia Statistics program *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) (lib.stat.cmu.edu/DASL/), 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. Paul is the father of two boys.

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 and his wife relax by biking or hiking, spending much of their free time in Canada, the Rockies, or the Blue Ridge Mountains. They have a son, a daughter, and four grandchildren.

Table of Contents

Part VII Inference When Variables Are Related

*Optional Chapter

Preface

We are often asked why we write Statistics texts. After all, it takes a lot of work to find new and better examples, to keep datasets current, and to make a book an enjoyable and effective learning tool. So we thought we'd address that question first.

We do it because it's fun.

Of course, we care about teaching students to think statistically; we are teachers and professional statisticians. But Statistics can be a particularly challenging subject to teach. The student encounters many new concepts, many new methods, and many new terms. And we want to change the way students think about the world. From the start, our challenge has been to write a book that students would read, learn from, and enjoy. And we return to that goal with each new edition.

The book you hold is quicker to the point and, we hope, even more readable than previous editions. Of course, we've kept our conversational style and background anecdotes.[1] But we've tightened discussions and adjusted the order of some topics to move the story we tell about Statistics even more quickly to interesting real-world questions. We've focused even more on statistical thinking.

More and more high school math teachers are using examples from Statistics to provide intuitive examples of how a little bit of math can help us say a lot about the world. So students expect Statistics to be about real-world insights. This edition of *Stats: Data and Models* keeps your students engaged and interested because we show Statistics in action right from the start. Students will be solving problems of the kind they're likely to encounter in real life sooner. In Chapter 4, they will be comparing groups and in Chapter 6, they'll see relationships between two quantitative variables—and, of course, always with real, modern data. By emphasizing statistical thinking, these early chapters lay the foundation for the more advanced chapters on Analysis of Variance and Multiple Regression at the end of the book, telling a consistent story throughout.

There are few things more fun and useful for students than being empowered to discover something new about the world. And few things more fun for authors than helping students make those discoveries.

So, What's New in This Edition?

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. Many chapters lead with new examples—and we follow through with analyses of the data from those examples.

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.

1. New and improved pedagogical tools: A new section head list at the beginning of each chapter provides a road map. Section heads within each chapter are reorganized and reworded to be clear and specific. Chapter study materials now include Learning Objectives as well as terms. Students who understand the objectives and know the terms are well on their way to being ready for any test.

2. Streamlined design: Our goal has always been an accessible text. This edition sports an entirely new design that clarifies the purpose of each text element. The major theme of each chapter is more linear and easier to follow without distraction. Essential supporting material is clearly boxed and shaded, so students know where to focus their study efforts. Enriching—and often entertaining—side material is boxed, but not shaded.

3. Streamlined content: Our reorganization has shortened the book from 31 to 29 chapters (Chapter 29 is provided on DVD and also online at pearsonhighered.com/dvb). Each chapter is still a focused discussion, and most can be taught in one lesson. We've combined topics that are conceptually similar and reduced time spent on secondary topics. We've grouped important concepts, often in new presentation orders. The result is a more readable text.

4. Content changes: Here's how we've reorganized the content:

   a. Chapter 1 now gets down to business immediately rather than just providing an introduction to the book's features.

   b. The discussions of probability and random variables are reorganized to improve clarity and flow; they are tighter and more to the point.

   c. We've moved the discussion of inference for means earlier. We still lead the discussion of inference with inference for proportions (for reasons we explain in the *Test Bank and Resource Guide*), but now we turn immediately to inference for means so students can see the methods side by side. Students can then also see that the reasoning is really the same.

   d. When we discuss comparing groups, we now discuss both proportions and means, which helps students to see the parallels.

5. Exercises: We've updated most exercises that use real-world data, retired some that were getting old, and added new exercises. Each chapter's exercises now start with single-concept exercises for each

[1] And our footnotes.

section, labeled with the section number so students can find exercises to review for any topic they wish to check.

6. We've updated the choice of technologies supported in the **On the Computer** sections at the end of chapters. You'll now find advice on *StatCrunch* and *R* along with the other packages we have traditionally discussed.
7. Sections are numbered to help with navigation and reading assignments.

Our Approach

Statistics is practiced with technology. We think a modern statistics text should recognize that fact from the start. And so do our students. You won't find tedious calculations worked by hand. But you will find equation forms that favor intuition over calculation. You'll find extensive use of real data—even large data sets. And, most important, you'll find a focus on statistical thinking rather than calculation. The question that motivates each of our hundreds of examples is not "how do you *find* the answer?" but "how do you *think* about the answer?"

Textbooks are defined more by what they choose not to cover than by what they do cover. We've structured this text so that each new topic fits into a student's growing understanding. Several topic orders can support this goal. We explain our reasons for our topic order in the *Test Bank and Resource Guide*. We also describe some alternative orders supported by these materials.

GAISE Guidelines

The Guidelines for Assessment and Instruction in Statistics Education (GAISE) report adopted by the American Statistical Association 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've designed our text and supplementary materials to support this approach to the introductory course. We urge you to think about these guidelines with each class meeting.

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 this edition even more approachable:

- ***Readability.*** This book doesn't read like other Statistics texts. Our style is both colloquial and informative, engaging students to actually read the book to see what it says. We've tightened the discussions and removed digressions.
- ***Humor.*** We know that humor is the best way to promote learning. You will find quips and wry comments throughout the narrative, in margin notes, and in footnotes.
- ***Informality.*** Our informal diction doesn't mean that we treat the subject matter lightly or informally. We try to be precise and, wherever possible, we offer deeper (but not more mathematical) explanations and justifications than those found in most introductory texts.
- ***Focused lessons.*** The chapters are shorter than in most other texts so instructors and students can focus on one topic at a time.
- ***Consistency.*** We try to avoid the "do what we say, not what we do" trap. Having taught the importance of plotting data and checking assumptions and conditions, we model that behavior right through the rest of the book. (Check the exercises in Chapter 28. You'll see that we still require and demonstrate the plots and checks that were introduced in the early chapters.) This consistency helps reinforce these fundamental principles.
- ***The need to read.*** Students who just skim the book, or start from an exercise and look for a similar example in a box to copy, may find our presentation frustrating. The important concepts, definitions, and sample solutions don't sit in little boxes. Statistics is a consistent story about how to understand the world when we have data. The story can't be told piecemeal. This is a book that needs to be read, so we've tried to make the reading experience enjoyable.

Mathematics

Mathematics can

1. provide a concise, clear statement of important concepts.
2. describe calculations to be performed with data.
3. 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 don't shy away from using math where it can clarify without intimidating. But we know that some students

are put off by equations, so we always provide a verbal description and a numerical example as well. Some theorems about Statistics 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. So we avoid them here.

Nor do we slide in the opposite direction and concentrate on calculation. 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 spend time summing squared deviations by hand. We have selected the equations that do appear for their focus on illuminating concepts and methods. Although these equations may be the best way to understand the concepts, they may not be optimal for hand calculation. When that happens, we give an alternative formula, better suited for hand calculation, for those who find following the process a better way to learn about the result.

Technology and Data

To experience the real world of Statistics, use modern technology to explore real data sets.

Technology. We assume that you are using some form of technology—a statistics package, a calculator, a spreadsheet, or some combination of these—in your Statistics course. We also assume that you'll put little emphasis on calculating answers by hand, even though we often show how. However, this is not a technology-heavy book. The role of technology in this book is to get the calculations out of the way so we can focus on statistical thinking. We discuss generic computer output, but we don't adopt any particular statistics software. We do offer guidance to help students get started on eight common software platforms: Excel®, Minitab®, *Data Desk*, JMP®, SPSS®, TI-83/84 Plus graphing calculators, StatCrunch®, and R®. The **On the Computer** section at the end of most chapters is specific to the methods learned in that chapter. All of these packages have inexpensive student options. But your students can have a choice of three of them at *no-cost: StatCrunch* (accessed through *MyStatLab*, available from Pearson with the text), *Data Desk* (found on the DVD and website that accompanies the book and free to move to their computers), and *R* (found online, but run on their computers.) The book's DVD includes the e-book *ActivStats*; the statistics package *Data Desk*; as well as versions of the data sets used in the book appropriate for a variety of packages.

Data. Because we use technology for computing, we don't limit ourselves to small, artificial data sets. You'll find some small data sets, but we also base 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. Machine-readable versions of the data are included on the DVD and on the book's website, **www.pearsonhighered.com/dvb**.

Continuing Features

Enhancing Understanding

Where Are We Going? Each chapter starts with a paragraph that points out the kinds of questions students will learn how to answer in the chapter. A new chapter outline helps organize major topics for the students.

Each chapter ends with a **What Have We Learned?** summary, which includes new learning objectives and definitions of terms introduced in the chapter. Students can think of these as study guides.

In each chapter, our innovative **What Can Go Wrong?** sections highlight the most common errors that people make and the misconceptions they have about Statistics. One of our goals is to arm students with the tools to detect statistical errors and to offer practice in debunking misuses of statistics, whether intentional or not.

Margin and in-text boxed notes. Throughout each chapter, boxed margin and in-text notes enhance and enrich the text. Boxes with essential information are screened. Conversational notes that enhance the text and entertain the reader are unscreened.

By Hand. Even though we encourage the use of technology to calculate statistical quantities, we realize the pedagogical benefits of occasionally doing a calculation by hand. The By Hand boxes break apart the calculation of many simpler formulas to help the student through the calculation of a worked example.

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.

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 are clear about the letters and symbols statisticians use to mean very specific things, so we've included Notation Alerts whenever we introduce a special notation that students will see again.

Connections. Each chapter has a Connections feature to link key terms and concepts with previous discussions and to point out the continuing themes. Connections help students

fit newly learned concepts into a growing understanding of Statistics.

Learning by Example

For Example. As we introduce each important concept, we provide a focused example applying it—usually with real up-to-the-minute data. Many For Examples carry the discussion through the chapter, picking up the story and moving it forward as students learn more about the topic.

Step-by-Step Examples: Think, Show, Tell. Step-by-Step examples repeat the mantra of Think, Show, and Tell in every chapter. These longer, worked examples guide students through the process of analyzing the problem with the general explanation on the left and the worked-out problem on the right. 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. The result is a better understanding of the concept, not just number crunching. In the fourth edition, we've updated Think, Show, Tell Step-by-Step examples with new examples and data.

Testing Understanding

Just Checking. Just Checking questions are quick checks throughout the chapter; most involve very little calculation. These questions encourage students to pause and think about what they've just read. The Just Checking answers are at the end of the exercise sets in each chapter so students can easily check themselves.

Exercises. We've added section-specific single-concept exercises at the beginning of each exercise 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. Exercises have been updated with the most recent data. Many come from news stories; some from recent research articles. Whenever possible, the data are on the bound-in DVD so students can explore them further.

Technology

***ActivStats* Pointers.** The DVD bound into new copies of the book includes *ActivStats*, so we've included occasional pointers to the *ActivStats* activities when they parallel discussions in the book. Many students choose to look at these first, before reading the chapter or attending a class on each subject.

Data Sources. Most of the data used in examples and exercises are from real-world sources, and whenever we can, we include references to the Internet data sources used, often in the form of URLs. The data we use are usually on the DVD. If you seek the data—or an updated version of the data—on the Internet, we try to direct you to a good starting point.

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 most chapters, we summarize what students can find in the most common packages, often with annotated output. We then offer specific guidance for several of the most common packages (*Data Desk*, Excel®, JMP®, Minitab®, R®, SPSS®, StatCrunch®, and TI-83/84 Plus[2]) to help students get started with the software of their choice.

On the DVD

The DVD accompanying new books holds a number of supporting materials, including *ActivStats,* the *Data Desk* statistics package, animations, all large data sets from the text formatted for the most popular technologies, and one additional chapter.

***ActivStats* (for *Data Desk*).** The award-winning *ActivStats* multimedia program supports learning chapter by chapter. It complements the book with videos of real-world 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. *ActivStats* includes

- more than 1000 homework exercises, plus answers to the "odd numbered" exercises.
- 17 short video clips, 70 animated activities, 117 teaching applets, and more than 300 data sets.

Data Desk, a student version of the professional statistics package, gives students the ability to perform any analysis in the textbook. Its interactive dynamic displays help students visualize relationships and models.

Data. Data for exercises marked are available on the DVD, Companion website, and MyStatLab™ formatted for multiple statistics software applications.

Additional Chapter. An additional chapter covering more advanced topics in multiple regression (Chapter 29). This chapter discusses modern diagnostic and model building methods.

[2]For brevity, we will write TI-83/84 Plus for the TI-83 Plus and/or TI-84 Plus. Keystrokes and output remain the same for the TI-83 Plus and the TI-84 Plus, so instructions and examples serve for both calculators.

Supplements

For the Student

***Stats: Data and Models*, Fourth Edition,** for-sale student edition (ISBN-13: 978-0-321-98649-8; ISBN-10: 0-321-98649-0)

Student's Solutions Manual, by William Craine, provides detailed, worked-out solutions to odd-numbered exercises. (ISBN-13: 978-0-321-98997-0; ISBN-10: 0-321-98997-X)

Study card for the De Veaux/Velleman/Bock Statistics Series 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: 978-0-321-82626-8; ISBN-10: 0-321-82626-4)

Videos for the De Veaux/Velleman/Bock Series, Fourth Edition, available to download from MyStatLab®. Concept Videos use humor to promote learning. Unique characters in fun situations explain the key concepts of statistics, covering important definitions and procedures for most chapters. Also available are videos of worked solutions for many of the Step-by-Step examples in the text.

For the Instructor

Instructor's Edition contains answers to all exercises. (ISBN-13: 978-0-321-99028-0; ISBN-10: 0-321-99028-5)

Instructor website contains the following resources at: **www.pearsonhighered.com/dvb**

- **Getting Started:** The De Veaux/Velleman/Bock Approach, Sample syllabi, Getting Started with Technology.
- **Preparing for Class:** Chapter and Lesson Support: ActivStats Pointers, Planning Your Lessons, Instructor's Supplements.

Instructor's Solutions Manual (download only), by William Craine, contains detailed solutions to all of the exercises. The Instructor's Solutions Manual is available to download from within MyStatLab® and in the Instructor Resource Center at **www.pearsonhighered.com/irc**.

Online Test Bank and Resource Guide (download only), by William Craine, includes chapter-by-chapter comments on the major concepts, tips on presenting topics (and what to avoid), extra teaching examples, a list of resources, chapter quizzes, part-level tests, and suggestions for projects. The Online Test Bank and Resource Guide is available to download from within MyStatLab® and in the Instructor Resource Center at **www.pearsonhighered.com/irc**.

Instructor's Podcasts (10 points in 10 minutes). These audio podcasts focus on key points in each chapter to help you with class preparation. They can be easily downloaded from MyStatLab and the Instructor Resource Center (**www.pearsonhighered.com/irc**).

Technology Resources

MyStatLab™ Online Course (access code required)

MyStatLab from Pearson is the world's leading online resource for teaching and learning statistics; integrating interactive homework, assessment, and media in a flexible, easy-to-use format. MyStatLab is a course management system that delivers **proven results** in helping individual students succeed.

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

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

- **Personalized Learning:** We now offer your course with an optional focus on adaptive learning, to allow your students to work on just what they need to learn when it makes the most sense, to maximize their potential for understanding and success.
- **Tutorial Exercises with Multimedia Learning Aids:** The homework and practice exercises in MyStatLab align with the exercises in the textbook, and most regenerate algorithmically to give students unlimited opportunity for practice and mastery. Exercises offer immediate helpful feedback, guided solutions, sample problems, animations, videos, statistical software tutorial videos and eText clips for extra help at point-of-use.
- **MyStatLab Accessibility:** MyStatLab is compatible with the JAWS screen reader, and enables multiple-choice and free-response problem-types to be read, and interacted with via keyboard controls and math notation input. MyStatLab also works with screen enlargers, including ZoomText, MAGic, and SuperNova. And all MyStatLab videos accompanying texts with copyright 2009 and later have closed captioning. More information on this functionality is available at http://mymathlab.com/accessibility.
- **StatTalk Videos:** Fun-loving statistician Andrew Vickers takes to the streets of Brooklyn, NY, to demonstrate important statistical concepts through interesting stories and real-life events. This series of 24 fun and engaging videos will help students actually understand statistical concepts. Available with an instructor's user guide and assessment questions.

- **Additional Question Libraries:** In addition to algorithmically regenerated questions that are aligned with your textbook, MyStatLab courses come with two additional question libraries.
  - **450 exercises** in **Getting Ready for Statistics** cover the developmental math topics students need for the course. These can be assigned as a prerequisite to other assignments, if desired.
  - **1000 exercises** in the **Conceptual Question Library** require students to apply their statistical understanding.
- **StatCrunch™:** MyStatLab integrates the web-based statistical software, StatCrunch, within the online assessment platform so that students can easily analyze data sets from exercises and the text. In addition, MyStatLab includes access to **www.StatCrunch.com**, a vibrant online community where users can access tens of thousands of shared data sets, create and conduct online surveys, perform complex analyses using the powerful statistical software, and generate compelling reports. Designed for today's students, StatCrunch works on any mobile device.
- **Statistical Software Support and Integration:** We make it easy to copy our data sets, both from the eText and the MyStatLab questions, into software such as StatCrunch, Minitab, Excel, and more. Students have access to a variety of support tools—Technology Tutorial Videos, Technology Study Cards, and Technology Manuals for select titles—to learn how to effectively use statistical software.

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

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

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

MyStatLab™ Ready to Go Course (access code required)

These new Ready to Go courses provide students with all the same great MyStatLab features, but make it easier for instructors to get started. Each course includes pre-assigned homework and quizzes to make creating a course even simpler. Ask your Pearson representative about the details for this particular course or to see a copy of this course.

MathXL® for Statistics Online Course (access code required)

MathXL® is the homework and assessment engine that runs MyStatLab. (MyStatLab is MathXL plus a learning management system.)

With MathXL for Statistics, instructors can:

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

With MathXL for Statistics, students can:

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

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

StatCrunch™

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

- **Collect**. Users can upload their own data to StatCrunch or search a large library of publicly shared data sets, spanning almost any topic of interest. Also, an online survey tool allows users to quickly collect data via web-based surveys.
- **Crunch**. A full range of numerical and graphical methods allow users to analyze and gain insights from any data set. Interactive graphics help users understand statistical concepts, and are available for export to enrich reports with visual representations of data.
- **Communicate.** Reporting options help users create a wide variety of visually-appealing representations of their data.

Full access to StatCrunch is available with a MyStatLab kit, and StatCrunch is available by itself to qualified adopters. StatCrunch Mobile works on any mobile device and is now available, visit www.statcrunch.com from the browser on your smart phone or tablet. For more information, visit our website at www.StatCrunch.com, or contact your Pearson representative.

TestGen®

TestGen® (www.pearsoned.com/testgen) enables instructors to build, edit, print, and administer tests using a computerized bank of questions developed to cover all the objectives of the text. TestGen is algorithmically based, allowing instructors to create multiple but equivalent versions of the same question or test with the click of a button. Instructors can also modify test bank questions or add new questions. The software and testbank are available for download from Pearson Education's online catalog.

PowerPoint® Lecture Slides

PowerPoint® Lecture Slides provide an outline to use in a lecture setting, presenting definitions, key concepts, and figures from the text. These slides are available within MyStatLab and in the Instructor Resource Center at **www.pearsonhighered.com/irc**.

Active Learning Questions

Prepared in PowerPoint®, these questions are intended for use with classroom response systems. Several multiple-choice questions are available for each chapter of the book, allowing instructors to quickly assess mastery of material in class. The Active Learning Questions are available to download from within MyStatLab® and in the Instructor Resource Center at **www.pearsonhighered.com/irc**.

Acknowledgments

Many people have contributed to this book in all of its editions. This edition never would have seen the light of day without the assistance of the incredible team at Pearson. Our Editor-in-Chief, Deirdre Lynch, was central to the genesis, development, and realization of this project from day one. Shannon Steed, Project Manager, kept the cogs from getting into the wheels where they often wanted to wander with much needed humor and grace. Senior Marketing Manager Erin Kelly made sure the word got out. Chere Bemelmans, Program Manager, Salena Casha and Justin's, Editorial Assistants, were essential in managing all of the behind-the-scenes work that needed to be done. Stephanie Green, Media Producer, put together a top-notch media package for this book. Melissa Welch of Studio Montage and Barbara Atkinson are responsible for the wonderful way the book looks. Carol Melville, Procurement Specialist, worked miracles to get this book in your hands, and Greg Tobin, President, was supportive and good-humored throughout all aspects of the project.

A special thanks goes out to Nancy Kincade, Project Manager at Lumina Datamatics, for her close attention to detail.

We are grateful for expert help from William Craine who developed and wrote supplements to accompany this new edition.

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 Cathleen Zucco-Tevloff, Rider University; and Stanley Seltzer, Ithaca College. Special thanks to Kurt Mederer, University of Michigan, for his careful reading of the text.

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

Nazanin Azarnia
Santa Fe Community College

Patricia Humphrey
Georgia Southern University

Steve Marsden
Glendale College

Cathy Zucco-Teveloff
Rider University

Jay Xu
Williams College

Ken Grace
Anoka Ramsey Community College

Joseph Kupresanin
Cecil College

Jackie Miller
The Ohio State University

Dottie Walton
Cuyahoga Community College

We also extend our sincere thanks for the suggestions and contributions made by the following reviewers of the previous editions:

Mary Kay Abbey
Montgomery College

Froozan Pourboghnaf Afiat
Community College of Southern Nevada

Mehdi Afiat
Community College of Southern Nevada

Nazanin Azarnia
Santa Fe Community College

Sanjib Basu
Northern Illinois University

Carl D. Bodenschatz
University of Pittsburgh

Steven Bogart
Shoreline Community College

Ann Cannon
Cornell College

Robert L. Carson
Hagerstown Community College

Jerry Chen
Suffolk County CommunityCollege

Rick Denman
Southwestern University

Jeffrey Eldridge
Edmonds Community College

Karen Estes
St. Petersburg Junior College

Richard Friary
Kim (Robinson) Gilbert
Clayton College & State University

Ken Grace
Anoka-Ramsey Community College

Jonathan Graham
University of Montana

Nancy Heckman
University of British Columbia

James Helreich
Marist College

Susan Herring
Sonoma State University

Mary R. Hudachek-Buswell
Clayton State University

Patricia Humphrey
Georgia Southern University

Becky Hurley
Rockingham CommunityCollege

Debra Ingram
Arkansas State University

Kelly Jackson
Camden County College

Martin Jones
College of Charleston

Rebecka Jornsten
Rutgers University

Michael Kinter
Cuesta College

Kathleen Kone
Community College of Allegheny County

Michael Lichter
State University of New York–Buffalo

Susan Loch
University of Minnesota

Pamela Lockwood
Western Texas A & M University

Wei-Yin Loh
University of Wisconsin–Madison

Catherine Matos
Clayton College & State University

Elaine McDonald
Sonoma State University

Hari Mukerjee
Witchita State University

Helen Noble
San Diego State University

Monica Oabos
Santa Barbara City College

Linda Obeid
Reedley College

Charles C. Okeke
Community College of Southern Nevada

Pamela Omer
Western New England College

Mavis Pararai
Indiana University of Pennsylvania

Gina Reed
Gainesville College

Juana Sanchez
UCLA

Gerald Schoultz
Grand Valley State University

Jim Smart
Tallahassee Community College

Chamont Wang
The College of New Jersey

Edward Welsh
Westfield State College

Heydar Zahedani
California State University, San Marcos

chapter 1

Stats Starts Here[1]

Where are we going?

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.

This is a book about understanding the world by using data. So we'd better start by understanding data. There's more to that than you might have thought.

"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

1.1 What Is Statistics?

People around the world have one thing in common—they all want to figure out what's going on. You'd think with the amount of information available to everyone today this would be an easy task, but actually, as the amount of information grows, so does our need to understand what it can tell us.

At the base of all this information, on the Internet and all around us, are data. We'll talk about data in more detail in the next section, but for now, think of **data** as any collection of numbers, characters, images, or other items that provide information about something. What sense can we make of all this data? You certainly can't make a coherent picture from random pieces of information. Whenever there are data and a need for understanding the world, you'll find Statistics.

This book will help you develop the skills you need to understand and communicate the knowledge that can be learned from data. By thinking clearly about the question you're trying to answer and learning the statistical tools to show what the data are saying, you'll acquire the skills to tell clearly what it all means. Our job is to help you make sense of the concepts and methods of Statistics and to turn it into a powerful, effective approach to understanding the world through data.

[1]We were thinking of calling this chapter "Introduction" but nobody reads the introduction, and we wanted you to read this. We feel safe admitting this down here in the footnotes because nobody reads footnotes either.

"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, former Director of Data Mining and Personalization, Amazon.com

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.

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."

Data vary. Ask different people the same question and you'll get a variety of answers. Statistics helps us to make sense of the world described by our data by seeing past the underlying variation to find patterns and relationships. This book will teach you skills to help with this task and ways of thinking about variation that are the foundation of sound reasoning about data.

Consider the following:

- If you have a Facebook account, you have probably noticed that the ads you see online tend to match your interests and activities. Coincidence? Hardly. According to the *Wall Street Journal* (10/18/2010),[2] much of your personal information has probably been sold to marketing or tracking companies. Why would Facebook give you a free account and let you upload as much as you want to its site? Because your data are valuable! Using your Facebook profile, a company might build a profile of your interests and activities: what movies and sports you like; your age, sex, education level, and hobbies; where you live; and, of course, who your friends are and what *they* like. From Facebook's point of view, your data are a potential gold mine. Gold ore in the ground is neither very useful nor pretty. But with skill, it can be turned into something both beautiful and valuable. What we're going to talk about in this book is how you can mine your own data and learn valuable insights about the world.
- In 2012, in the United States, wireless subscribers (about 90% of the U.S. population) sent a total of 2.19 trillion text (SMS) messages. That's over 182 billion a month, up from about 12 million a month just 10 years earlier.[3] Some of these messages are sent or read while the sender or the receiver is driving. How dangerous is texting while driving?

  How can we study the effect of texting while driving? One way is to measure reaction times of drivers faced with an unexpected event while driving and texting. Researchers at the University of Utah tested drivers on simulators that could present emergency situations. They compared reaction times of sober drivers, drunk drivers, and texting drivers.[4] The results were striking. The texting drivers actually responded more slowly and were more dangerous than those who were above the legal limit for alcohol.

In this book, you'll learn how to design and analyze experiments like this. You'll learn how to interpret data and to communicate the message you see to others. You'll also learn how to spot deficiencies and weaknesses in conclusions drawn by others that you see in newspapers and on the Internet every day. Statistics can help you become a more informed citizen by giving you the tools to understand, question, and interpret data.

[2] blogs.wsj.com/digits/2010/10/18/referers-how-facebook-apps-leak-user-ids/

[3] CTIA The International Association for the Wireless Telecommunications Industry (www.ctia.org/your-wireless-life/how-wireless-works/wireless-quick-facts).

[4] "Text Messaging During Simulated Driving," Drews, F. A. et al., Human Factors: hfs.sagepub.com/content/51/5/762

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).*

Paleontology: *Previous Life.*

Biology: *Life.*

Religion: *After Life*

Anthropology: *Who?*

History: *What, where, and when?*

Philosophy: *Why?*

Engineering: *How?*

Accounting: *How much?*

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

1.2 Data

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 2012, the company's sales reached $61 billion (a 27% increase from the previous year). Amazon has sold a wide variety of merchandise, including a $400,000 necklace, yak cheese from Tibet, and the largest book in the world. How did Amazon become so successful and how can it keep track of so many customers and such a wide variety of products? The answer to both questions is *data.*

But what are data? 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 your name and address in Amazon's database are also data even though they are not numerical. What about your ZIP code? That's a number, but would Amazon care about, say, the *average* ZIP code of its customers?

Let's look at some hypothetical values that Amazon might collect:

105-2686834-3759466	Ohio	Nashville	Kansas	10.99	440	N	B00000I5Y6	Katherine H.
105-9318443-4200264	Illinois	Orange County	Boston	16.99	312	Y	B000002BK9	Samuel P.
105-1872500-0198646	Massachusetts	Bad Blood	Chicago	15.98	413	N	B000068ZVQ	Chris G.
103-2628345-9238664	Canada	Let Go	Mammals	11.99	902	N	B0000010AA	Monique D.
002-1663369-6638649	Ohio	Best of Kansas	Kansas	10.99	440	N	B002MXA7Q0	Katherine H.

Try to guess what they represent. Why is that hard? Because there is no *context*. If we don't know what values are measured and what is measured about them, the values are meaningless. We can make the meaning clear if we organize the values into a **data table** such as this one:

Order Number	Name	State/Country	Price	Area Code	Previous Album Download	Gift?	ASIN	New Purchase Artist
105-2686834-3759466	Katherine H.	Ohio	10.99	440	Nashville	N	B0000015Y6	Kansas
105-9318443-4200264	Samuel R	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
105-1372500-0198646	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
103-2628345-9238664	Monique D.	Canada	11.99	902	Let Go	N	B0000010AA	Mammals
002-1663369-6638649	Katherine H.	Ohio	10.99	440	Best of Kansas	N	B002MXA7Q0	Kansas

Now we can see that these are purchase records for album download orders from Amazon. The column titles tell what has been recorded. Each row is about a particular purchase.

What information would provide a **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. The answers to the first two questions are essential. If we don't know *what* values are measured and *who* those values are measured on, the values are meaningless.

Who and What

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. Cases go by different names, depending on the situation.

- Individuals who answer a survey are called **respondents**.
- People on whom we experiment are **subjects** or (in an attempt to acknowledge the importance of their role in the experiment) **participants**.
- Animals, plants, websites, and other inanimate subjects are often called **experimental units**.
- Often we simply call cases what they are: for example, *customers, economic quarters*, or *companies*.
- In a database, rows are called **records**—in this example, purchase records. Perhaps the most generic term is *cases*, but in any event the rows represent the *who* of the data.

Data trumps intuition Amazon monitors and evolves its website to better serve customers and maximize sales. To decide which changes to make, analysts experiment with new designs, offers, recommendations, and links. Statisticians want to know how long you'll spend browsing the site and whether you'll follow the links or purchase the suggested items. As Ronny Kohavi, former director of Data Mining and Personalization for Amazon, 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."

Look at all the columns to see exactly what each row refers to. Here the cases are different purchase records. You might have thought that each customer was a case, but notice that, for example, Katherine H. appears twice, both in the first and the last row. A common place to find out exactly what each row refers to is the leftmost column. That value often identifies the cases, in this example, it's the order number. If you collect the data yourself, you'll know what the cases are. But, often, you'll be looking at data that someone else collected and you'll have to ask or figure that out yourself.

Often the cases are a **sample** from some larger **population** that we'd like to understand. Amazon doesn't care about just these customers; it wants to understand the buying patterns of *all* its customers, and, generalizing further, it wants to know how to attract other Internet users who may not 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.

We must know *who* and *what* to analyze data. Without knowing these two, we don't have enough information 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. And knowing *why* the data were collected can tell us much about its reliability and quality.

How the Data Are Collected

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. Identifying the W's is a habit we recommend.

The first step of any data analysis is to know what you are trying to accomplish and what you want to know. To help you use Statistics to understand the world and make decisions, we'll lead you through the entire process of *thinking* about the problem, *showing* what you've found, and *telling* others what you've learned. Every guided example in this book is broken into these three steps: *Think*, *Show,* and *Tell*. Identifying the problem and the *who* and *what* of the data is a key part of the *Think* step of any analysis. Make sure you know these before you proceed to *Show* or *Tell* anything about the data.

A S ***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.

For Example IDENTIFYING THE "WHO"

In December 2013, *Consumer Reports* published an evaluation of 46 tablets from a variety of manufacturers.

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

ANSWER: The magazine is interested in the performance of tablets currently offered for sale. It tested a sample of 46 tablets, which are the "Who" for these data. Each tablet selected represents all similar tablets offered by that manufacturer.

1.3 Variables

The characteristics recorded about each individual are called **variables**. They are usually found as the columns of a data table with a name in the header that identifies what has been recorded. In the Amazon data table we find the variables *Order Number, Name, State/Country, Price,* and so on.

Categorical Variables

Some variables just tell us what group or category each individual belongs to. Are you male or female? Pierced or not? We call variables like these **categorical**, or **qualitative, variables**. (You may also see them called **nominal variables** because they name categories.) Some variables are clearly categorical, like the variable *State/Country*. Its values are text and those values tell us what category the particular case falls into. But numerals are often used to label categories, so categorical variable values can also be numerals. For example, Amazon collects telephone area codes that *categorize* each phone number into a geographical region. So area code is considered a categorical variable even though it has numeric values. (But see the story in the following box.)

Area codes—numbers or categories? 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. Since the advent of push-button phones, area codes have finally become just categories.

"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"**

Descriptive responses to questions are often categories. For example, the responses to the questions "Who is your cell phone provider?" or "What is your marital status?" yield categorical values. When Amazon considers a special offer of free shipping to customers, it might first analyze how purchases have been shipped in the recent past. Amazon might start by counting the number of purchases shipped in each category: ground transportation, second-day air, and overnight air. Counting is a natural way to summarize a categorical variable like *Shipping Method*. Chapter 2 discusses summaries and displays of categorical variables more fully.

Quantitative Variables

When a variable contains measured numerical values with measurement *units*, we call it a **quantitative variable**. Quantitative variables typically record an amount or degree of something. For quantitative variables, its measurement **units** provide a meaning for the numbers. Even more important, units such as yen, cubits, carats, angstroms, nanoseconds,

miles per hour, or degrees Celsius tell us the *scale* of measurement, so we know 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, pennies, yen, or Estonian krooni. Chapter 3 discusses quantitative variables. We'll see how to display and summarize them there.

Sometimes a variable with numeric values can be treated as either categorical or quantitative depending on what we want to know from it. Amazon could record your *Age* in years. That seems quantitative, and it would be if the company wanted to know the average age of those customers who visit their site after 3 AM. But suppose Amazon wants to decide which album to feature on its site when you visit. Then thinking of your age in one of the categories Child, Teen, Adult, or Senior might be more useful. So, sometimes whether a variable is treated as categorical or quantitative is more about the question we want to ask rather than an intrinsic property of the variable itself.

Identifiers

For a categorical variable like *Sex*, each individual is assigned one of two possible values, say *M* or *F*. But for a variable with ID numbers, such as a *student ID*, each individual receives a unique value. We call a variable like this, which has exactly as many values as cases, an **identifier variable**. Identifiers are useful, but not typically for analysis.

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. Amazon also wants to send you the right product, so it assigns a unique Amazon Standard Identification Number (ASIN) to each item it carries. You'll want to recognize when a variable is playing the role of an identifier so you aren't tempted to analyze it.

Identifier variables themselves don't tell us anything useful about their categories because we know there is exactly one individual in each. However, they are crucial in this era of large data sets because by uniquely identifying the cases, they make it possible to combine data from different sources, protect (or violate) privacy, and provide unique labels. Many large databases are *relational* databases. In a relational database, different data tables link to one another by matching identifiers. In the Amazon example, the *Customer Number*, *ASIN*, and *Transaction Number* are all identifiers. The IP (Internet protocol) address of your computer is another identifier, needed so that the electronic messages sent to you can find you.

Privacy and the Internet You have many Identifiers: a social security number, a student ID number, possibly a passport number, a health insurance number, and probably a Facebook account name. Privacy experts are worried that Internet thieves may match your identity in these different areas of your life, allowing, for example, your health, education, and financial records to be merged. Even online companies such as Facebook and Google are able to link your online behavior to some of these identifiers, which carries with it both advantages and dangers. The National Strategy for Trusted Identities in Cyberspace (www.wired.com/images_blogs/threatlevel/2011/04/NSTICstrategy_041511.pdf) proposes ways that we may address this challenge in the near future.

Ordinal Variables

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? Often the best way to tell is to look to the *why* of the study. 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 purely quantitative. To treat it as quantitative, she'll have to imagine that it has "educational value units" or some similar arbitrary construct. Because there are no natural units, she should be cautious. Variables 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 and understand what you want to learn from the variable to decide whether to treat it as categorical or quantitative.

For Example IDENTIFYING "WHAT" AND "WHY" OF TABLETS

RECAP: A *Consumer Reports* article about 46 tablets lists each tablet's manufacturer, price, battery life (hrs.), the operating system (Android, iOS, or Windows), an overall quality score (0–100), and whether or not it has a memory card reader.

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

ANSWER: The variables are

- manufacturer (categorical)
- price (quantitative, $)
- battery life (quantitative, hrs.)
- operating system (categorical)
- performance score (quantitative, no units)
- memory card reader (categorical)

The magazine hopes to provide consumers with the information to choose a good tablet.

Just Checking

In the 2004 Tour de France, Lance Armstrong made history by winning the race for an unprecedented sixth time. In 2005, he became the only 7-time winner and set a new record for the fastest average speed—41.65 kilometers per hour—that stands to this day. In 2012, he was banned for life for doping offenses and stripped of all of his titles. You can find data on all the Tour de France races in the data set **Tour de France 2014**. Here are the first three and last eight lines of the data set. Keep in mind that the entire data set has over 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	Age	Team	Total Time (h/min/s)	Avg. Speed (km/h)	Stages	Total Distance Ridden (km)	Starting Riders	Finishing Riders
1903	Maurice Garin	France	32	La Française	94.33.00	25.7	6	2428	60	21
1904	Henri Cornet	France	20	Cycles JC	96.05.00	25.3	6	2428	88	23
1905	Louis Trousseller	France	24	Peugeot	112.18.09	27.1	11	2994	60	24
⋮										
2007	Alberto Contador	Spain	24	Discovery Channel	91.00.26	38.97	20	3547	189	141
2008	Carlos Sastre	Spain	33	CSC-Saxo Bank	87.52.52	40.50	21	3559	199	145
2009	Alberto Contador	Spain	26	Astana	85.48.35	40.32	21	3460	180	156
2010	Andy Schleck	Luxembourg	25	Saxo Bank	91.59.27	39.590	20	3642	180	170
2011	Cadel Evans	Australia	34	BMC	86.12.22	39.788	21	3430	198	167
2012	Bradley Wiggins	Great Britain	32	Sky	87.34.47	39.827	20	3488	198	153
2013	Christopher Froome	Great Britain	28	Sky	83.56.40	40.551	21	3404	198	169
2014	Vincenzo Nibali	Italy	29	Astana	89.56.06	40.735	21	3663.5	198	164

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 website. 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 extra symbols such 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 data and what they represent.** The same variable can sometimes take on different roles.
- **Don't assume that a variable is quantitative just because its values are numbers.** 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 website. The question that respondents answered may be posed in a way that influences responses.

What Have We Learned?

Learning Objectives

Understand that data are values, whether numerical or labels, together with their context.

- *Who, what, why, where, when* (and *how*)—the W's—help nail down the context of the data.
- We must know *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.
- Stop and identify the W's whenever you have data, and be sure you can identify the cases and the variables.

Consider the source of your data and the reasons the data were collected. That can help you understand what you might be able to learn from the data.

Identify whether a variable is being used as categorical or quantitative.

- Categorical variables identify a category for each case. Usually we think about the counts of cases that fall in 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 may treat the same variable as categorical or quantitative depending on what we want to learn from it, which means some variables can't be pigeonholed as one type or the other.

Review of Terms

The key terms are in chapter order so you can use this list to review the material in the chapter.

Data Recorded values whether numbers or labels, together with their context (p. 1).

Data table An arrangement of data in which each row represents a case and each column represents a variable (p. 3).

Context The context ideally tells *who* was measured, *what* was measured, *how* the data were collected, *where* the data were collected, and *when* and *why* the study was performed (p. 4).

Case An individual about whom or which we have data (p. 4).

Respondent Someone who answers, or responds to, a survey (p. 4).

Subject A human experimental unit. Also called a participant (p. 4).

Participant A human experimental unit. Also called a subject (p. 4).

Experimental unit An individual in a study for which or for whom data values are recorded. Human experimental units are usually called subjects or participants (p. 4).

Record Information about an individual in a database (p. 4).

Sample A subset of a population, examined in hope of learning about the population (p. 4).

Population The entire group of individuals or instances about whom we hope to learn (p. 4).

Variable A variable holds information about the same characteristic for many cases (p. 5).

Categorical (or qualitative) variable A variable that names categories with words or numerals (p. 5).

Nominal variable The term "nominal" can be applied to a variable whose values are used only to name categories (p. 5).

Quantitative variable A variable in which the numbers are values of measured quantities with units (p. 5).

Units A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams (p. 5).

Identifier variable A categorical variable that records a unique value for each case, used to name or identify it (p. 6).

Ordinal variable The term "ordinal" can be applied to a variable whose categorical values possess some kind of order (p. 6).