

GLOBAL
EDITION



Stats

Data and Models

FOURTH EDITION

De Veaux • Velleman • Bock



ALWAYS LEARNING

PEARSON

Global 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*
Acquisitions Editor, Global Editions: *Murchana Borthakur*
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*
Project Editor, Global Editions: *K.K. Neelakantan*
Senior Production Manufacturing Controller, Global Editions:
Trudy Kimber
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*
Media Production Manager, Global Editions: *Vikram Kumar*
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: *Lumina Datamatics, Inc.*

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the world

Visit us on the World Wide Web at: www.pearsonglobaleditions.com

© Pearson Education Limited 2016

The rights of Richard D. De Veaux, Paul F. Velleman, and David E. Bock to be identified as the authors of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

Authorized adaptation from the United States edition, entitled Stats: Data and Models, 4th edition, ISBN 978-0-321-98649-8, by Ricgard D. De Veaux, Paul F. Velleman, and David E. Bock published by Pearson Education © 2016.

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 either the prior written permission of the publisher or a license permitting restricted copying in the United Kingdom issued by the Copyright Licensing Agency Ltd, Saffron House, 6–10 Kirby Street, London EC1N 8TS.

All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

10 9 8 7 6 5 4 3 2 1

ISBN 10: 1-292-10163-6
ISBN 13: 978-1-292-10163-7

Typeset by Lumina Datamatics, Inc.

Printed in Malaysia

*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

Preface 9

Part I Exploring and Understanding Data

1 Stats Starts Here 17

1.1 What Is Statistics? ■ 1.2 Data ■ 1.3 Variables

2 Displaying and Describing Categorical Data 32

2.1 Summarizing and Displaying a Single Categorical Variable ■ 2.2 Exploring the Relationship Between Two Categorical Variables

3 Displaying and Summarizing Quantitative Data 61

3.1 Displaying Quantitative Variables ■ 3.2 Shape ■ 3.3 Center ■
3.4 Spread ■ 3.5 Boxplots and 5-Number Summaries ■ 3.6 The Center of Symmetric Distributions: The Mean ■ 3.7 The Spread of Symmetric Distributions: The Standard Deviation ■ 3.8 Summary—What to *Tell* About a Quantitative Variable

4 Understanding and Comparing Distributions 100

4.1 Comparing Groups with Histograms ■ 4.2 Comparing Groups with Boxplots ■
4.3 Outliers ■ 4.4 Timeplots: Order, Please! ■ 4.5 Re-Expressing Data: A First Look

5 The Standard Deviation as a Ruler and the Normal Model 128

5.1 Standardizing with z-Scores ■ 5.2 Shifting and Scaling ■ 5.3 Normal Models ■ 5.4 Finding Normal Percentiles ■ 5.5 Normal Probability Plots

Part II Exploring Relationships Between Variables

6 Scatterplots, Association, and Correlation 167

6.1 Scatterplots ■ 6.2 Correlation ■ 6.3 Warning: Correlation \neq Causation ■
6.4 Straightening Scatterplots

7 Linear Regression 198

7.1 Least Squares: The Line of “Best Fit” ■ 7.2 The Linear Model ■ 7.3 Finding the Least Squares Line ■ 7.4 Regression to the Mean ■ 7.5 Examining the Residuals ■ 7.6 R^2 —The Variation Accounted For by the Model ■ 7.7 Regression Assumptions and Conditions

8 Regression Wisdom 235

8.1 Examining Residuals ■ 8.2 Extrapolation: Reaching Beyond the Data ■
8.3 Outliers, Leverage, and Influence ■ 8.4 Lurking Variables and Causation ■
8.5 Working with Summary Values

9 Re-expressing Data: Get It Straight! 263

9.1 Straightening Scatterplots – The Four Goals ■ 9.2 Finding a Good Re-Expression

Part III Gathering Data

10 Understanding Randomness 296

10.1 What Is Randomness? ■ 10.2 Simulating by Hand

11 Sample Surveys 310

11.1 The Three Big Ideas of Sampling ■ 11.2 Populations and Parameters ■
11.3 Simple Random Samples ■ 11.4 Other Sampling Designs ■ 11.5 From the
Population to the Sample: You Can't Always Get What You Want ■ 11.6 The Valid
Survey ■ 11.7 Common Sampling Mistakes, or How to Sample Badly

12 Experiments and Observational Studies 334

12.1 Observational Studies ■ 12.2 Randomized, Comparative Experiments ■
12.3 The Four Principles of Experimental Design ■ 12.4 Control Treatments ■
12.5 Blocking ■ 12.6 Confounding

Part IV Randomness and Probability

13 From Randomness to Probability 364

13.1 Random Phenomena ■ 13.2 Modeling Probability ■ 13.3 Formal Probability

14 Probability Rules! 382

14.1 The General Addition Rule ■ 14.2 Conditional Probability and the General
Multiplication Rule ■ 14.3 Independence ■ 14.4 Picturing Probability: Tables,
Venn Diagrams, and Trees ■ 14.5 Reversing the Conditioning and Bayes' Rule

15 Random Variables 405

15.1 Center: The Expected Value ■ 15.2 Spread: The Standard Deviation ■
15.3 Shifting and Combining Random Variables ■ 15.4 Continuous Random
Variables

16 Probability Models 428

16.1 Bernoulli Trials ■ 16.2 The Geometric Model ■ 16.3 The Binomial Model
■ 16.4 Approximating the Binomial with a Normal Model ■ 16.5 The Continuity
Correction ■ 16.6 The Poisson Model ■ 16.7 Other Continuous Random Variables:
The Uniform and the Exponential

Part V From the Data at Hand to the World at Large

17 Sampling Distribution Models 459

17.1 Sampling Distribution of a Proportion ■ 17.2 When Does the Normal Model
Work? Assumptions and Conditions ■ 17.3 The Sampling Distribution of Other
Statistics ■ 17.4 The Central Limit Theorem: The Fundamental Theorem of
Statistics ■ 17.5 Sampling Distributions: A Summary

18 Confidence Intervals for Proportions	488
18.1 A Confidence Interval ■ 18.2 Interpreting Confidence Intervals: What Does 95% Confidence Really Mean? ■ 18.3 Margin of Error: Certainty vs. Precision ■ 18.4 Assumptions and Conditions	
19 Testing Hypotheses About Proportions	510
19.1 Hypotheses ■ 19.2 P-Values ■ 19.3 The Reasoning of Hypothesis Testing ■ 19.4 Alternative Alternatives ■ 19.5 P-Values and Decisions: What to Tell About a Hypothesis Test	
20 Inferences About Means	534
20.1 Getting Started: The Central Limit Theorem (Again) ■ 20.2 Gosset's t ■ 20.3 Interpreting Confidence Intervals ■ 20.4 A Hypothesis Test for the Mean ■ 20.5 Choosing the Sample Size	
21 More About Tests and Intervals	564
21.1 Choosing Hypotheses ■ 21.2 How to Think About P-Values ■ 21.3 Alpha Levels ■ 21.4 Critical Values for Hypothesis Tests ■ 21.5 Errors	

Part VI Accessing Associations Between Variables

22 Comparing Groups	601
22.1 The Standard Deviation of a Difference ■ 22.2 Assumptions and Conditions for Comparing Proportions ■ 22.3 A Confidence Interval for the Difference Between Two Proportions ■ 22.4 The Two Sample z -Test: Testing for the Difference Between Proportions ■ 22.5 A Confidence Interval for the Difference Between Two Means ■ 22.6 The Two-Sample t -Test: Testing for the Difference Between Two Means ■ 22.7 The Pooled t -Test: Everyone into the Pool?	
23 Paired Samples and Blocks	646
23.1 Paired Data ■ 23.2 Assumptions and Conditions ■ 23.3 Confidence Intervals for Matched Pairs ■ 23.4 Blocking	
24 Comparing Counts	671
24.1 Goodness-of-Fit Tests ■ 24.2 Chi-Square Test of Homogeneity ■ 24.3 Examining the Residuals ■ 24.4 Chi-Square Test of Independence	
25 Inferences for Regression	705
25.1 The Population and the Sample ■ 25.2 Assumptions and Conditions ■ 25.3 Intuition About Regression Inference ■ 25.4 Regression Inference ■ 25.5 Standard Errors for Predicted Values ■ 25.6 Confidence Intervals for Predicted Values ■ 25.7 Logistic Regression	

Part VII Inference When Variables Are Related

*26 Analysis of Variance	763
26.1 Testing Whether the Means of Several Groups Are Equal ■ 26.2 The ANOVA Table ■ 26.3 Assumptions and Conditions ■ 26.4 Comparing Means ■ 26.5 ANOVA on Observational Data	
27 Multifactor Analysis of Variance	798
27.1 A Two Factor ANOVA Model ■ 27.2 Assumptions and Conditions ■ 27.3 Interactions	
28 Multiple Regression	833
28.1 What Is Multiple Regression? ■ 28.2 Interpreting Multiple Regression Coefficients ■ 28.3 The Multiple Regression Model—Assumptions and Conditions ■ 28.4 Multiple Regression Inference ■ 28.5 Comparing Multiple Regression Models	
29 Multiple Regression Wisdom (available online)	875
29.1 Indicators ■ 29.2 Diagnosing Regression Models: Looking at the Cases ■ 29.3 Building Multiple Regression Models	
Appendixes	
A Answers A-1 ■ B Photo Acknowledgments A-51 ■ C Index A-53 ■ D Tables and Selected Formulas A-69	

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 online at www.pearsonglobaleditions.com/deveaux). 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

¹And 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 website that accompanies the book and free to move to their computers), and *R* (found online, but run on their computers.) The book's website 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 book's website, www.pearsonglobaleditions.com/deveaux.

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 provided so students can explore them further.

Technology

ActivStats Pointers. The website 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 website. 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 Plus²) to help students get started with the software of their choice.

²For 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

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.

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 website contains the following resources at: www.pearsonglobaleditions.com/deveaux

- **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.pearsonglobaleditions.com/deveaux.

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.pearsonglobaleditions.com/deveaux.

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.pearsonglobaleditions.com/deveaux).

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.

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.pearsonglobaleditions.com/deveaux.

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.pearsonglobaleditions.com/deveaux.

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 Inc., 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 <i>Santa Fe Community College</i>	Ken Grace <i>Anoka Ramsey Community College</i>
Patricia Humphrey <i>Georgia Southern University</i>	Joseph Kupresanin <i>Cecil College</i>
Steve Marsden <i>Glendale College</i>	Jackie Miller <i>The Ohio State University</i>
Cathy Zucco-Teveloff <i>Rider University</i>	Dottie Walton <i>Cuyahoga Community College</i>
Jay Xu <i>Williams College</i>	

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

Mary Kay Abbey <i>Montgomery College</i>	Nazanin Azarnia <i>Santa Fe Community College</i>
Froozan Pourboghna Afiat <i>Community College of Southern Nevada</i>	Sanjib Basu <i>Northern Illinois University</i>
Mehdi Afiat <i>Community College of Southern Nevada</i>	Carl D. Bodenschatz <i>University of Pittsburgh</i>
	Steven Bogart <i>Shoreline Community College</i>

Ann Cannon <i>Cornell College</i>	Michael Lichter <i>State University of New York–Buffalo</i>
Robert L. Carson <i>Hagerstown Community College</i>	Susan Loch <i>University of Minnesota</i>
Jerry Chen <i>Suffolk County Community College</i>	Pamela Lockwood <i>Western Texas A & M University</i>
Rick Denman <i>Southwestern University</i>	Wei-Yin Loh <i>University of Wisconsin–Madison</i>
Jeffrey Eldridge <i>Edmonds Community College</i>	Catherine Matos <i>Clayton College & State University</i>
Karen Estes <i>St. Petersburg Junior College</i>	Elaine McDonald <i>Sonoma State University</i>
Richard Friary Kim (Robinson) Gilbert <i>Clayton College & State University</i>	Hari Mukerjee <i>Wichita State University</i>
Ken Grace <i>Anoka Ramsey Community College</i>	Helen Noble <i>San Diego State University</i>
Jonathan Graham <i>University of Montana</i>	Monica Oabos <i>Santa Barbara City College</i>
Nancy Heckman <i>University of British Columbia</i>	Linda Obeid <i>Reedley College</i>
James Helreich <i>Marist College</i>	Charles C. Okeke <i>Community College of Southern Nevada</i>
Susan Herring <i>Sonoma State University</i>	Pamela Omer <i>Western New England College</i>
Mary R. Hudachek-Buswell <i>Clayton State University</i>	Mavis Pararai <i>Indiana University of Pennsylvania</i>
Patricia Humphrey <i>Georgia Southern University</i>	Gina Reed <i>Gainesville College</i>
Becky Hurley <i>Rockingham Community College</i>	Juana Sanchez <i>UCLA</i>
Debra Ingram <i>Arkansas State University</i>	Gerald Schoultz <i>Grand Valley State University</i>
Kelly Jackson <i>Camden County College</i>	Jim Smart <i>Tallahassee Community College</i>
Martin Jones <i>College of Charleston</i>	Chamont Wang <i>The College of New Jersey</i>
Rebecka Jornsten <i>Rutgers University</i>	Edward Welsh <i>Westfield State College</i>
Michael Kinter <i>Cuesta College</i>	Heydar Zahedani <i>California State University, San Marcos</i>
Kathleen Kone <i>Community College of Allegheny County</i>	

- 1.1 What Is Statistics?
- 1.2 Data
- 1.3 Variables

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.

¹We 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.



FRAZZ © 2003 Jef Mallett. Distributed by Universal Uclick. Reprinted with permission. All rights reserved.

“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.

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

³CTIA 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	B0000015Y6	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.

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.

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.”

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.

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.

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.

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. 17).
Data table	An arrangement of data in which each row represents a case and each column represents a variable (p. 19).
Context	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 (p. 20).
Case	An individual about whom or which we have data (p. 20).
Respondent	Someone who answers, or responds to, a survey (p. 20).
Subject	A human experimental unit. Also called a participant (p. 20).
Participant	A human experimental unit. Also called a subject (p. 20).
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. 20).
Record	Information about an individual in a database (p. 20).
Sample	A subset of a population, examined in hope of learning about the population (p. 20).
Population	The entire group of individuals or instances about whom we hope to learn (p. 20).
Variable	A variable holds information about the same characteristic for many cases (p. 21).
Categorical (or qualitative) variable	A variable that names categories with words or numerals (p. 21).
Nominal variable	The term "nominal" can be applied to a variable whose values are used only to name categories (p. 21).
Quantitative variable	A variable in which the numbers are values of measured quantities with units (p. 21).
Units	A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams (p. 21).
Identifier variable	A categorical variable that records a unique value for each case, used to name or identify it (p. 22).
Ordinal variable	The term "ordinal" can be applied to a variable whose categorical values possess some kind of order (p. 22).

On the Computer: ENTERING DATA

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.

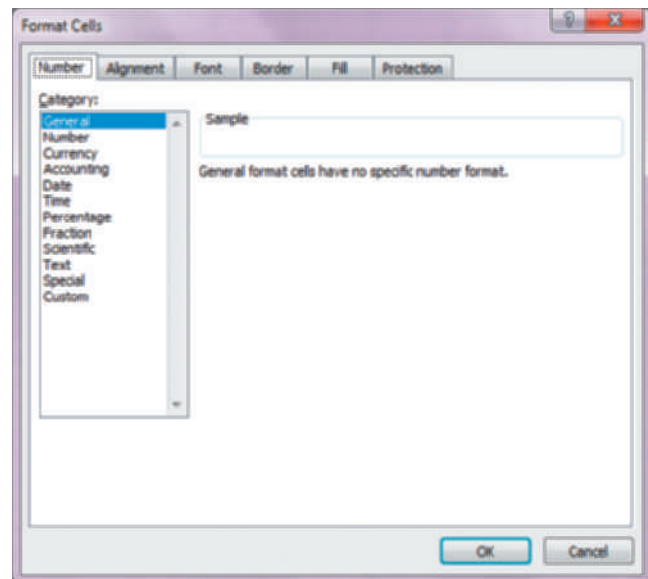
For example, to get your data into a computer statistics package, you need to tell the computer:

- 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 with cases in the rows and variables in the columns. Most computer statistics packages prefer the *delimiter* that marks the division between elements of a data table to be a tab character (comma is another common delimiter) 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.
- Excel is often used to help organize, manipulate and prepare data for other software packages. Many of the other packages take Excel files as inputs. Alternatively, you can copy a data table from Excel and Paste it into many packages, or export Excel spreadsheets as tab delimited (.txt) or comma delimited files (.csv), which can be easily shared and imported into other programs. All data files provided with this text are in these two formats.

EXCEL

To open a file containing data in Excel:

- Choose **File > Open**.
- Browse to find the file to open. Excel supports many file formats.
- Other programs can import data from a variety of file formats, but all take both tab delimited (.txt) and comma delimited (.csv) text files.
- You can also copy tables of data from other sources, such as Internet sites, and paste them into an Excel spreadsheet. Excel can recognize the format of many tables copied this way, but this method may not work for some tables.
- Excel may not recognize the format of the data. If data include dates or other special formats (\$, €, ¥, etc.), identify the desired format. Select the cells or columns to reformat and choose **Format > Cell**. Often, the General format is the best option.



DATA DESK

To read data into Data Desk

- **File > Open**. The dialog lets you specify variable names (or take them from the first row of the data), the delimiter, or how to read formatted data.
- The **File > Import** command works in the same way, but instead of starting a new data file, it adds the data in the file to the current data file. Data Desk can work with multiple data tables in the same file.
- If the data are already in another program, such as, for example, a spreadsheet, **Copy** the data table (including the column headings). In Data Desk choose **Edit > Paste Variables**. There is no need to create variables first; Data Desk does that automatically. You'll see the same dialog as for Open and Import.

JMP

To import a text file, choose:

- **File > Open** and select the file from the dialog. At the bottom of the dialog screen you'll see **Open As:**—be sure to change to **Data (Using Preview)**. This will allow you to specify the delimiter and make sure the variable names are correct. (**JMP** also allows various formats to be imported directly, including.xls files.)

You can also paste a data set in directly (with or without variable names) by selecting:

- **File > New > New Data Table** and then **Edit > Paste (or Paste with Column Names** if you copied the names of the variables as well).

Finally, you can import a data set from a URL directly by selecting:

- **File > Internet Open** and pasting in the address of the web site. JMP will attempt to find data on the page. It may take a few tries and some edits to get the data set in correctly.

MINITAB

To import a text or Excel file, choose:

- **File > Open Worksheet.** From **Files of type,** choose **Text (*.txt)** or **Excel (*.xls; *xlsx).**
- Browse to find and select the file.

- In the lower right corner of the dialog, choose **Open** to open the data file alone, or **Merge** to add the data to an existing worksheet.
- Click **Open.**

R

R can import many types of files, but text files (tab or comma delimited) are easiest. If the file is tab delimited and contains the variable names in the first row, then:

```
> mydata = read.delim(file.choose())
```

will give a dialog where you can pick the file you want to import. It will then be in a data frame called mydata. If the file is comma delimited, use:

```
>mydata=read.csv(file.choose())
```

COMMENTS

(RStudio provides an interactive dialog that may be easier to use). For other options including the case that the file does not contain variable names, consult **R** help.

SPSS

To import a text file, choose:

- **File > Open > Data.** Under "Files of type," choose **Text (*.txt,*.dat).** Select the file you want to import. Click **Open.**

- A window will open called **Text Import Wizard.** Follow the steps, depending on the type of file you want to import.

STATCRUNCH

Statcrunch offers several ways to enter data. Click MyStatCrunch > My Data. Click a dataset to analyze the data or edit its properties.

MY DATA

Click a data set link to analyze the data or edit its properties. Want to load a new data set?

- Select a file on my computer.
- Enter the WWW address of a file.

- Paste data into a form.
- Type or paste data into a blank data table.

For the "select a file on my computer" option, Statcrunch offers a choice of whitespace, comma, tab or semicolon delimiters. You may also choose to use the first line as the names of the variables.

After making your choices, select the **Load File** button at the bottom of the screen.

TI 83/84 PLUS

Locate the TI List on the book's website pearsonglobaleditions.com.

- Link your calculator to the computer using the USB cable.
- Right-click on the file name and choose "Send to TI-Device."

Exercises

Section 1.1

1. **Grocery shopping** Many grocery store chains offer customers a card they can scan when they check out and offer discounts to people who do so. To get the card, customers must give information, including a mailing address and e-mail address. The actual purpose is not to reward loyal customers but to gather data. What data do these cards allow stores to gather, and why would they want that data?
2. **Online shopping** Online retailers such as Amazon.com keep data on products that customers buy, and even products they look at. What does Amazon hope to gain from such information?

Section 1.2

3. **Super Bowl** Sports announcers love to quote statistics. During the Super Bowl, they particularly love to announce when a record has been broken. They might have a list of all Super Bowl games, along with the scores of each team, total scores for the two teams, margin of victory, passing yards for the quarterbacks, and many more bits of information. Identify the *who* in this list.
4. **Nobel laureates** The website www.nobelprize.org allows you to look up all the Nobel prizes awarded in any year. The data are not listed in a table. Rather you drag a slider to the year and see a list of the awardees for that year. Describe the *who* in this scenario.

Section 1.3

5. **Grade levels** A person's grade in school is generally identified by a number.
 - a) Give an example of a *why* in which grade level is treated as categorical.
 - b) Give an example of a *why* in which grade level is treated as quantitative.

6. **ZIP codes** The Postal Service uses five-digit ZIP codes to identify locations to assist in delivering mail.
 - a) In what sense are ZIP codes categorical?
 - b) Is there any ordinal sense to ZIP codes? In other words, does a larger ZIP code tell you anything about a location compared to a smaller ZIP code?
7. **Voters** A February 2010 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?
8. **Job hunting** A June 2011 Gallup Poll asked Americans, "Thinking about the job situation in America today, would you say that it is now a good time or a bad time to find a quality job?" The choices were "Good time" or "Bad time." What kind of variable is the response?
9. **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?
10. **Stress** A medical researcher measures the increase in heart rate of patients under a stress test. What kind of variable is the researcher studying?

Chapter Exercises

11. **The news** Find a newspaper or magazine article in which some data are reported. For the data discussed in the article, identify as many of the W's as you can. Include a copy of the article with your report.
12. **The Internet** Find an Internet source that reports on a study and describes the data. Print out the description and identify as many of the W's as you can.

(Exercises 13–20) For each description of data, identify Who and What were investigated and the Population of interest.

13. **Height?** A study conducted by a team of researchers found that during pregnancy, a woman can tell whether a man is short or tall by looking at his face. To explore the subject, the authors conducted three investigations, the first of which involved 40 undergraduate women who were asked to guess the likely height of 80 men based on photos of their face. Half of the men were short, and the other half were tall. All held similar expressions in the photos or were deemed to be equally attractive. None of the women were in their third trimester at the time of the test. The result: the closer a woman was to her third trimester the more accurate her guess.
 14. **Hula-hoops** The hula-hoop, a popular children’s toy in the 1950s, has gained popularity as an exercise in recent years. But does it work? To answer this question, the American Council on Exercise conducted a study to evaluate the cardio and calorie-burning benefits of “hooping.” Researchers recorded heart rate and oxygen consumption of participants, as well as their individual ratings of perceived exertion, at regular intervals during a 30-minute workout. (www.acefitness.org/certifiednewsarticle/1094/)
 15. **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 3500 cars passing him, he found that when he wore his helmet, motorists passed 3.29 inches closer to him, on average, than when his head was bare. (Source: *NY Times*, Dec. 10, 2006)
 16. **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.
 17. **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. (Source: *NY Times*, Dec. 10, 2006)
 18. **Blindness** A study begun in 2011 examines the use of stem cells in treating two forms of osteopenia. Each of the 24 patients entered one of two separate trials in which embryonic stem cells were to be used to treat the condition.
 19. **Candy bars** A look at 515 participants found that participants who ate two or more candy bars a day “experienced waist size increases six times greater than those of people who didn’t eat the candy bars.”
 20. **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 21–34) 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).*
21. **Weighing hummingbirds** Because of the difficulty of weighing a hummingbird in the mountains, researchers caught and measured 34 hummingbirds, recording their weight, wingspan, body length, and sex. They hoped to find a way to estimate weight from the other, more easily determined quantities.
 22. **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.
 23. **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.
 24. **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.

- 25. Training** Researchers investigating the impact of training on sprinting collected data from 668 training sessions during 1998–2000. They kept track of the runner’s age, the number of weeks the training lasted, the type of training (interval, endurance, combined), the level of training the runner had (none, moderate, intense), the weight and sex of the runners, and whether the runners exhibited improvements (none, minor, major).
- 26. 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.
- 27. 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 benefits of the compound.
- 28. Orchards** Business analysts hoping to provide information helpful to American cherry growers compiled these data about cherry trees: size (square kilometers), number of months in existence, state, varieties of cherries grown, average price per ton, gross sales, and percent profit.
- 29. 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).
- 30. 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.
- 31. Refrigerators** In 2013, *Consumer Reports* published an article evaluating refrigerators. It listed 353 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).
- 32. 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. (Source: *STATS* No. 39, Winter 2004)
- 33. Kentucky Derby 2014** 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.

Year	Winner	Jockey	Trainer	Owner	Time
2013	Orb	J. Rosario	S. McGaughey	Stuart Janney & Phipps Stable	2:02.89
2012	I’ll Have Another	M. Gutierrez	D. O’Neill	Reddam Racing	2:01.83
2011	Animal Kingdom	J. Velazquez	H. G. Motion	Team Valor	2:02.04
2010	Super Saver	C. Borel	T. Pletcher	WinStar Farm	2:04.45
2009	Mine That Bird	C. Borel	B. Woolley	Double Eagle Ranch	2:02.66
2008	Big Brown	K. Desormeaux	R. Dutrow	IEAH Stables, Pompa et al	2:01.82
...					
1878	Day Star	J. Carter	L. Paul	T. J. Nichols	2:37.25
1877	Baden Baden	W. Walker	E. Brown	Daniel Swigert	2:38
1876	Vagrant	R. Swim	J. Williams	William Astor	2:38.25
1875	Aristides	O. Lewis	A. Williams	H. P. McGrath	2:37.75

Source: Excerpt from HorseHats.com. Published by Thoroughbred Promotions.