

SIXTH EDITION

Intro Stats

MPH
17.4

BPM
154

RPM
79

MILES
21.3

FEET
262

CALORIES
1143

TIME
01:14:29



De Veaux | Velleman | Bock

This page intentionally left blank

SIXTH EDITION

Intro Stats

Richard D. De Veaux

Williams College

Paul F. Velleman

Cornell University

David E. Bock

Ithaca High School (Retired)

with contributions from

Brianna Heggeseth

Macalaster College

and

Susan Wang

Google Inc.

Content Development: Robert Carroll

Content Management: Suzanna Bainbridge

Content Production: Rose Kernan (RPK Editorial Services), Rachel S. Reeve, Nicolas Sweeny

Product Management: Karen Montgomery

Product Marketing: Alicia Wilson

Rights and Permissions: Tanvi Bhatia, Rimpay Sharma

Cover Credits: Photo: MeskPhotography/Shutterstock, City map: Ex_artist/Shutterstock, Mountain icon: Martial Red/Shutterstock

Please contact <https://support.pearson.com/getsupport/s/> with any queries on this content

Microsoft and/or its respective suppliers make no representations about the suitability of the information contained in the documents and related graphics published as part of the services for any purpose. All such documents and related graphics are provided “as is” without warranty of any kind. Microsoft and/or its respective suppliers hereby disclaim all warranties and conditions with regard to this information, including all warranties and conditions of merchantability, whether express, implied or statutory, fitness for a particular purpose, title and non-infringement. In no event shall Microsoft and/or its respective suppliers be liable for any special, indirect or consequential damages or any damages whatsoever resulting from loss of use, data or profits, whether in an action of contract, negligence or other tortious action, arising out of or in connection with the use or performance of information available from the services.

The documents and related graphics contained herein could include technical inaccuracies or typographical errors. Changes are periodically added to the information herein. Microsoft and/or its respective suppliers may make improvements and/or changes in the product(s) and/or the program(s) described herein at any time. Partial screen shots may be viewed in full within the software version specified.

Microsoft® and Windows® are registered trademarks of the Microsoft Corporation in the U.S.A. and other countries. This book is not sponsored or endorsed by or affiliated with the Microsoft Corporation.

Copyright © 2022, 2018, 2014 by Pearson Education, Inc. or its affiliates, 221 River Street, Hoboken, NJ 07030. All Rights Reserved. Manufactured in the United States of America. This publication is protected by copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise. For information regarding permissions, request forms, and the appropriate contacts within the Pearson Education Global Rights and Permissions department, please visit www.pearsoned.com/permissions/.

Cover Images Credits: Flat map by Ex_artist/Shutterstock; Speed on bike by MeskPhotography/Shutterstock; Mountain peak top flag by Martial Red/Shutterstock

Acknowledgments of third-party content appear on page A-47, which constitutes an extension of this copyright page.

PEARSON, ALWAYS LEARNING, and MYLAB are exclusive trademarks owned by Pearson Education, Inc. or its affiliates in the U.S. and/or other countries.

Unless otherwise indicated herein, any third-party trademarks, logos, or icons that may appear in this work are the property of their respective owners, and any references to third-party trademarks, logos, icons, or other trade dress are for demonstrative or descriptive purposes only. Such references are not intended to imply any sponsorship, endorsement, authorization, or promotion of Pearson’s products by the owners of such marks, or any relationship between the owner and Pearson Education, Inc., or its affiliates, authors, licensees, or distributors.

Library of Congress Cataloging-in-Publication Data

Names: De Veaux, Richard D., author. | Velleman, Paul F., 1949- author. | Bock, David E., author.

Title: Intro stats / Richard D. De Veaux, Williams College, Paul F. Velleman, Cornell University, David E. Bock, Cornell University ; with contributions from Brianna Heggeseth, Macalaster College, and Susan Wang, Google Inc.

Description: Sixth edition. | Hoboken, NJ : Pearson, [2022] | Includes index. | Summary: “An introduction to Statistics using the authors’ signature tools for teaching about randomness, sampling distribution models, and interference”-- Provided by publisher.

Identifiers: LCCN 2021002385 | ISBN 9780136806868 (hardcover)

Subjects: LCSH: Statistics--Textbooks.

Classification: LCC QA276.12 .D4 2022 | DDC 519.5--dc23

LC record available at <https://lcn.loc.gov/2021002385>

ScoutAutomatedPrintCode



Access Code Card

ISBN-10: 0-13-680689-9

ISBN-13: 978-0-13-680689-9

Student Rental

ISBN-10: 0-13-680686-4

ISBN-13: 978-0-13-680686-8

*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.” He is the C. Carlisle and M. Tippit Professor of Statistics at Williams College, where he has taught since 1994. 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 and was the 2018–2021 Vice-President of the ASA. Dick is also well known in industry, where for more than 30 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 and with the Choeur Vittoria of Paris. 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*, the Internet site Data and Story Library (DASL) (DASL.datadesk.com), which provides data sets for teaching Statistics (and is one source for the datasets used in this text.), and the tools referenced in the text for simulation and bootstrapping. Paul’s understanding of using and teaching with technology informs much of this book’s approach.

Paul taught Statistics at Cornell University, where he was awarded the MacIntyre Award for Exemplary Teaching. He is Emeritus Professor of Statistical Science from Cornell and lives in Maine with his wife, Sue Michlovitz. 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. In his spare time he sings with the *a cappella* group VoXX and studies tai chi.



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 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.

Preface ix

Index of Applications xxi

PART I Exploring and Understanding Data

1 Stats Starts Here 1

1.1 What Is Statistics? ♦ 1.2 Data ♦ 1.3 Variables ♦ 1.4 Models

2 Displaying and Describing Data 18

2.1 Summarizing and Displaying a Categorical Variable ♦ 2.2 Displaying a Quantitative Variable ♦ 2.3 Shape ♦ 2.4 Center ♦ 2.5 Spread

3 Relationships Between Categorical Variables—Contingency Tables 67

3.1 Contingency Tables ♦ 3.2 Conditional Distributions ♦ 3.3 Displaying Contingency Tables ♦ 3.4 Three Categorical Variables

4 Understanding and Comparing Distributions 98

4.1 Displays for Comparing Groups ♦ 4.2 Outliers ♦ 4.3 Re-Expressing Data: A First Look

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

5.1 Using the Standard Deviation to Standardize Values ♦ 5.2 Shifting and Scaling ♦ 5.3 Normal Models ♦ 5.4 Working with Normal Percentiles ♦ 5.5 Normal Probability Plots

Review of Part I: Exploring and Understanding Data 163

PART II Exploring Relationships Between Variables

6 Scatterplots, Association, and Correlation 173

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

7 Linear Regression 207

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

*Indicates optional sections.

8 Regression Wisdom 247

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 ♦ ***8.6** Straightening Scatterplots—The Three Goals ♦ ***8.7** Finding a Good Re-Expression

9 Multiple Regression 292

9.1 What Is Multiple Regression? ♦ **9.2** Interpreting Multiple Regression Coefficients ♦ **9.3** The Multiple Regression Model—Assumptions and Conditions ♦ **9.4** Partial Regression Plots ♦ ***9.5** Indicator Variables

Review of Part II: Exploring Relationships Between Variables 323**PART III Gathering Data****10 Sample Surveys 335**

10.1 The Three Big Ideas of Sampling ♦ **10.2** Populations and Parameters ♦ **10.3** Simple Random Samples ♦ **10.4** Other Sampling Designs ♦ **10.5** From the Population to the Sample: You Can't Always Get What You Want ♦ **10.6** The Valid Survey ♦ **10.7** Common Sampling Mistakes, or How to Sample Badly

11 Experiments and Observational Studies 361

11.1 Observational Studies ♦ **11.2** Randomized, Comparative Experiments ♦ **11.3** The Four Principles of Experimental Design ♦ **11.4** Control Groups ♦ **11.5** Blocking ♦ **11.6** Confounding

Review of Part III: Gathering Data 385**PART IV From the Data at Hand to the World at Large****12 From Randomness to Probability 391**

12.1 Random Phenomena ♦ **12.2** Modeling Probability ♦ **12.3** Formal Probability ♦ **12.4** Conditional Probability and the General Multiplication Rule ♦ **12.5** Independence ♦ **12.6** Picturing Probability: Tables, Venn Diagrams, and Trees ♦ **12.7** Reversing the Conditioning and Bayes' Rule

12A: Random Variables and Probability Models (Online)

12A.1 Expected Value: Center ♦ **12A.2** Standard Deviation ♦ **12A.3** Combining Random Variables ♦ **12A.4** The Binomial Model ♦ **12A.5** Modeling the Binomial with a Normal Model ♦ ***12A.6** The Poisson Model ♦ **12A.7** Continuous Random Variables

13 Sampling Distribution Models and Confidence Intervals for Proportions 431

13.1 The Sampling Distribution Model for a Proportion ♦ **13.2** When Does the Normal Model Work? Assumptions and Conditions ♦ **13.3** A Confidence Interval for a Proportion ♦ **13.4** Interpreting Confidence Intervals: What Does 95% Confidence Really Mean? ♦ **13.5** Margin of Error: Certainty vs. Precision ♦ ***13.6** Choosing the Sample Size

14 Confidence Intervals for Means 464

14.1 The Central Limit Theorem ♦ **14.2** A Confidence Interval for the Mean
 ♦ **14.3** Interpreting Confidence Intervals ♦ ***14.4** Picking Our Interval Up by Our
 Bootstraps ♦ **14.5** Thoughts About Confidence Intervals

15 Testing Hypotheses 499

15.1 Hypotheses ♦ **15.2** P-Values ♦ **15.3** The Reasoning of Hypothesis Testing
 ♦ **15.4** A Hypothesis Test for the Mean ♦ **15.5** Intervals and Tests ♦ **15.6** P-Values and
 Decisions: What to Tell About a Hypothesis Test

16 More About Tests and Intervals 536

16.1 Interpreting P-Values ♦ **16.2** Alpha Levels and Critical Values ♦ **16.3** Practical vs.
 Statistical Significance ♦ **16.4** Errors

Review of Part IV: From the Data at Hand to the World at Large 563**PART V Inference for Relationships****17 Comparing Groups 571**

17.1 A Confidence Interval for the Difference Between Two Proportions ♦ **17.2** Assumptions
 and Conditions for Comparing Proportions ♦ **17.3** The Two-Sample z-Test: Testing the
 Difference Between Proportions ♦ **17.4** A Confidence Interval for the Difference Between
 Two Means ♦ **17.5** The Two-Sample *t*-Test: Testing for the Difference Between Two Means
 ♦ ***17.6** Pooling ♦ ***17.7** The Standard Deviation of a Difference

18 Paired Samples and Blocks 617

18.1 Paired Data ♦ **18.2** The Paired *t*-Test ♦ **18.3** Confidence Intervals for Matched Pairs
 ♦ **18.4** Blocking

19 Comparing Counts 642

19.1 Goodness-of-Fit Tests ♦ **19.2** Chi-Square Test of Homogeneity ♦ **19.3** Examining the
 Residuals ♦ **19.4** Chi-Square Test of Independence

20 Inferences for Regression 675

20.1 The Regression Model ♦ **20.2** Assumptions and Conditions ♦ **20.3** Regression
 Inference and Intuition ♦ **20.4** The Regression Table ♦ **20.5** Multiple Regression
 Inference ♦ **20.6** Confidence and Prediction Intervals ♦ ***20.7** Logistic Regression
 ♦ ***20.8** More About Regression

Review of Part V: Inference for Relationships 721**Parts I–V Cumulative Review Exercises 734****Appendixes**

A Answers **A-1** ♦ **B** Credits **A-47** ♦ **C** Indexes **A-55**
 ♦ **D** Tables and Selected Formulas **A-67**

This page intentionally left blank

Intro Stats, sixth edition, has been especially exciting to develop. The book you hold has several innovations. Of course, we've kept our conversational style and anecdotes,¹ but we've enriched that material with greater use of our signature tools for teaching about randomness, sampling distribution models, and inference throughout the book. We've added current discussions of ethical issues to each chapter. Each chapter now ends with a student project suitable as a challenge for collaborative work. We've expanded discussions of models for data to include models with more than two variables. We've taken our inspiration both from our experience in the classroom and from the 2016 revision of the Guidelines for Assessment and Instruction in Statistics Education (GAISE) report adopted by the American Statistical Association. As a result, we increased the text's innovative uses of technology to encourage more statistical thinking, while maintaining its traditional core concepts and coverage. You'll notice that the order of topics is designed, to expand our attention beyond just one or two variables.

Innovations

Technology

The GAISE guidelines call on us to *Use technology to explore concepts and analyze data*. We emphatically agree. We think a modern statistics text should recognize from the start that statistics is practiced with technology. And so should our students. You won't find tedious calculations worked by hand. You *will* find equation forms that favor intuition over calculation. You'll find extensive use of real data—even large data sets. Throughout, 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 calculate the answer?” but “How do you think about the answer?”

For this edition of *Intro Stats* we've taken this principle still further. We have harnessed technology to develop simulation tools to improve the learning of two of the most difficult concepts in the introductory course: the idea of a sampling distribution and the reasoning of statistical inference.

Multivariable Thinking and Multiple Regression

GAISE's first guideline is to give students experience with multivariable thinking. The world is not univariate, and relationships are not limited to two variables. The fifth edition of *Intro Stats* introduced a third variable as early as Chapter 3's discussion of contingency tables and mosaic plots. The positive responses we've seen to this innovation, have led us to build on it. Following the discussion of correlation and regression as a tool (that is, without inference) in Chapters 6, 7, and 8, we introduce multiple regression in Chapter 9.

Multiple regression may be the most widely used statistical method, and it is certainly one that students need to understand. It is easy to perform multiple regressions with any statistics program, and the exercise of thinking about more than two variables is worth the effort. We've added new material about interpreting what regression models say. The effectiveness of multiple regression is immediately obvious and makes the reach and power of statistics clear. The use of real data underscores the universal applicability of these methods.

When we return to regression in Chapter 20 to discuss inference, we can deal with both simple and multiple regression models together. There is nothing different to discuss. (For this reason we set aside the F -test and adjusted R^2 . Students can add those later if

¹And footnotes

they need them.) This course is an *introduction* to statistics. It isn't necessary to learn *all* the details of the methods and models. But it is important to come away with a sense of the power and usefulness of statistics to solve real problems.

Innovative ways to teach the logic of statistical inference have received increasing attention. Among these are greater use of computer-based simulations and resampling methods (randomization tests and bootstrapping) to teach concepts of inference.



Bootstrap

The introduction to the new GAISE guidelines explicitly mentions the bootstrap method. The bootstrap is not as widely available or as widely understood as multiple regression. But it fits our presentation naturally. In this edition, we have expanded and made more extensive use of our innovative, **Random Matters** feature. Random Matters elements provide students with hands-on experience with randomness, randomization, and ways statistics can use randomness. In early chapters they draw small samples repeatedly from large populations to illustrate how the randomness introduced by sampling leads to both sampling distributions and statistical reasoning for inference. But what can we do when we have only one sample? The bootstrap provides a way to continue this line of thought, now by re-sampling from the sample at hand.

Bootstrapping provides an elegant way to simulate sampling distributions that we might not otherwise be able to see. And it does not require the assumption of Normality expected by Student's t -based methods. However, these methods are not as widely available or widely used in other disciplines, so they should not be the only—or even the principal—methods taught. They may be able to enhance student understanding, but instructors may wish to downplay them if that seems best for a class. We've placed these sections strategically so that instructors can choose the level that they are comfortable with and that works best with their course.

Real Data

GAISE recommends that instructors integrate real data with a context and purpose. More and more high school math teachers are using examples from statistics to demonstrate intuitively how a little bit of math can help us say a lot about the world. So our readers expect statistics to be about real-world insights. *Intro Stats* keeps readers engaged and interested because we show statistics in action right from the start. The exercises pose problems of the kind likely to be encountered in real life and propose ways to think about making inferences almost immediately—and, of course, always with real, up-to-date data.

Let us be clear. *Intro Stats* comes with an archive of over 300 datasets used throughout the book. The datasets are available online at the student resource site, in MyLab Statistics and at the free site [DASL.datadescription.com](https://dasl.datadescription.com). Examples that use these datasets cite them in the text. More than 700 of our exercises have a  tag next to them to indicate that the dataset referenced in the exercise is available electronically. The exercise title or a note provides the dataset title. Some exercises have a  tag to indicate that they call for the student to generate random samples or use randomization methods such as the bootstrap. Although we hope students will have access to computers, we provide ample exercises with full computer output for students to read, interpret, and explain. We encourage students to get the datasets and reproduce our examples using their statistics software, and some of the exercises require that.

Ethics

GAISE also calls for discussions of ethical issues. In this edition, new discussions of relevant ethical concerns are found in every chapter. We have chosen topics motivated by current events and issues students will know about. These elements are good fodder for classroom discussions.

For example, the discussion on p. 77 addresses the conflict between offering survey respondents the freedom to self-identify variables such as their gender, religion, race, or political position and the challenge that poses for data privacy. Combinations of responses can become so narrow that the individual's identity can be inadvertently exposed. A widely advertised “brain supplement” that claims to have laboratory-based proof of efficacy is the subject of the Ethics Matters element on p. 521. Their posted test results show that they were “p-hacking” to find significance where none existed in the original data. On p. 218, we discuss troubling fact that some of the founders of Statistics, including Galton, Pearson, and Fisher were proponents of eugenics.

Student Projects

Each chapter ends with a new student project. These can be the basis for more extensive investigations by students working on their own. But they also have enough “meat” to support team efforts. Most require that students use computers to gather or to analyze data. All expect the resulting product will include discussion and conclusions and thus be more than just some numbers or P-values.

Streamlined Content

Following the GAISE recommendations, we've streamlined several parts of the course: Introductory material is covered more rapidly. Today's students have seen a lot of statistics in their K–12 math courses and in their daily contact with online and print news sources. We still cover the topics to establish consistent terminology (such as the difference between a histogram and a bar chart). Chapter 2 does most of the work that previously took two chapters.

The discussion of random variables and probability distributions is shorter than in previous editions—again, a GAISE recommendation. Those are interesting topics, but they are not needed in this course. We leave them for a later course for those students who want to go further.

The Random Matters features show students that statistics vary from sample to sample, show them (empirical) sampling distributions, note the effect of sample size on the shape and variation of the sampling distribution of the mean, and suggest that it looks Normal. As a result, the discussion of the Central Limit Theorem is transformed from the most difficult one in the course to a relatively short discussion (“What you think is true about means really is true; there's this theorem.”) that can lead directly to the reasoning of confidence intervals.

Finally, introducing multiple regression doesn't really add much material to the lesson on inference for multiple regression because little is new.

GAISE 2016

As we've said, all of these enhancements follow the new Guidelines for Assessment and Instruction in Statistics Education (GAISE) 2016 report adopted by the American Statistical Association:

1. Teach statistical thinking.
 - ◆ Teach statistics as an investigative process of problem-solving and decision-making.
 - ◆ Give students experience with multivariable thinking.
2. Focus on conceptual understanding.
3. Integrate real data with a context and purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

The result is a course that is more aligned with the skills needed in the 21st century, one that focuses even more on statistical thinking and makes use of technology in innovative ways, while retaining core principles and topic coverage.

The challenge has been to use this modern point of view to improve learning without discarding what is valuable in the traditional introductory course. Many first statistics courses serve wide audiences of students who need these skills for their own work in disciplines where traditional statistical methods are, well, traditional. So we have not reduced our emphasis on the concepts and methods you expect to find in our texts.

Chapter Order

We've streamlined the presentation of basic topics that most students have already seen. Pie charts, bar charts, histograms, and summary statistics all appear in Chapter 2. Chapter 3 introduces contingency tables, and Chapter 4 discusses comparing distributions. Chapter 5 introduces the Normal model and the 68–95–99.7 Rule. The four chapters of Part II then explore linear relationships among quantitative variables—but here we introduce only the models and how they help us understand relationships. We leave the inference questions until later in the book. Part III discusses how data are gathered by survey and experiment.

In Part IV, Chapter 12 introduces basic probability and prepares us for inference. Naturally, a new approach to teaching inference has led to a reorganization of inference topics. In Chapter 13 we introduce confidence intervals for proportions as soon as we've reassured students that their intuition about the sampling distribution of proportions is correct. Chapter 14 formalizes the Central Limit Theorem and introduces Student's t models. Chapter 15 is then about testing hypotheses, and Chapter 16 elaborates further, discussing alpha levels, Type I and Type II errors, power, and effect size. The subsequent chapters in Part V deal with comparing groups (both with proportions and with means), paired samples, chi-square, and finally, inferences for regression models (both simple and multiple).

We've found that one of the challenges students face is how to know what technique to use when. In the real world, questions don't come at the ends of the chapters. So, as always, we've provided summaries at the end of each part along with a series of exercises designed to stretch student understanding. These Part Reviews are a mix of questions from all the chapters in that part. Finally, we've added an extra set of “book-level” review problems at the end of the text. These ask students to integrate what they've learned from the entire course. The questions range from simple questions about what method to use in various situations to more complete data analyses from real data. We hope that these will provide a useful way for students to organize their understanding at the end of the course.

Our Approach

We've discussed how this book is different, but there are some things we haven't changed.

- ◆ **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.
- ◆ **Humor.** 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 explanations and justifications than those found in most introductory texts.
- ◆ **Focused lessons.** The chapters are shorter than in most other texts so that 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 through the rest of the book. (Check out the exercises in Chapter 20.)

- ◆ **The need to read.** 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. Students who start with the exercises and then search back for a worked example that looks the same but with different numbers will find that our presentation doesn't support that approach.

Mathematics

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 discouraged by equations, so we always provide a verbal description and a numerical example as well.

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, today, virtually all statistics are calculated with technology. We have selected the equations that focus on illuminating concepts and methods rather than for hand calculation. We sometimes give an alternative formula, better suited for hand calculation, for those who find that following the calculation process is a better way to learn about the result.

Technology and Data

We assume that computers and appropriate software are available—at least for demonstration purposes. We hope that students have access to computers and statistics software for their analyses. We make more extensive use of special applications to demonstrate properties of randomness, illustrate the concept of a sampling distribution, and offer bootstrap methods for inference. These applications can be found in MyLab Statistics and at **astools.datadesk.com**.

We discuss generic computer output at the end of most chapters, but we don't adopt any particular statistics software. The **Tech Support** sections at the ends of chapters offer guidance for seven common software platforms: Data Desk, Excel, JMP, Minitab, SPSS, StatCrunch, and R. We also offer some advice for TI-83/84 Plus graphing calculators, although we hope that those who use them will also have some access to computers and statistics software.

We don't limit ourselves to small, artificial data sets, but base most examples and exercises on real data with a moderate number of cases. Machine-readable versions of the data are available at the Pearson Math & Stats Resource Site, MyLab Statistics, and at **dasl.datadescription.com**.

Features

Enhancing Understanding

Where Are We Going? Each chapter starts with a paragraph that raises the kinds of questions we deal with in the chapter. A chapter outline organizes the major topics and sections.

Random Matters. These innovative features travel along a progressive path of understanding randomness and our data. The first Random Matters element begins our thinking about drawing inferences from data. Subsequent Random Matters draw histograms of sample means, introduce the thinking involved in permutation tests, and encourage judgment about how likely the observed statistic seems when viewed against the simulated sampling distribution of the null hypothesis (without, of course, using those terms). Later Random Matters elements lead students through bootstrap calculations and compare

bootstrap results to classical inference. The Random Matters elements have been rewritten and expanded to provide step-by-step guidance.

New! Ethics Matters. This feature introduces relevant ethical considerations. Each chapter has an ethics discussion. These discuss current ethical issues. All are good material for classroom discussion, which we encourage.

New! Student projects. Each chapter ends with a student project that uses the methods learned thus far. They can be used for individual work or as a basis for team projects.

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.

Each chapter ends with several elements to help students study and consolidate what they've seen in the chapter.

- ◆ **Connections** specifically ties the new topics to those learned in previous chapters.
- ◆ **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.
- ◆ Next, the **Chapter Review** summarizes the story told by the chapter and provides a bullet list of the major concepts and principles covered.
- ◆ A **Review of Terms** is a glossary of all of the special terms introduced in the chapter. In the text, these are printed in **bold** and underlined. The Review provides page references, so students can easily turn back to a full discussion of the term if the brief definition isn't sufficient.

The **Tech Support** section provides the commands in each of the supported statistics packages that deal with the topic covered by the chapter. These are not full documentation, but should be enough to get a student started in the right direction.

Learning by Example



Step-by-Step Examples. We have updated the examples in our innovative Step-by-Step feature. Each one provides a longer, worked example that guides students through the process of analyzing a problem. The examples follow our three-step Think, Show, Tell organization for approaching a statistics task. They are organized with general explanations of each step on the left and a worked-out solution on the right. The right side of the grid models what would be an "A" level solution to the problem. Step-by-Steps illustrate 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. Our emphasis is on statistical thinking, and the pedagogical result is a better understanding of the concept, not just number crunching.

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

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.

Assessing Understanding

Our **Exercises** have some special features worth noting. In the initial exercises, you’ll find relatively simple, focused problems organized by chapter section. After that come more extensive exercises that may deal with topics from several parts of the chapter or even from previous chapters as they combine with the topics of the chapter at hand. All exercises appear in pairs. The odd-numbered exercises have answers in the back of student texts. Each even-numbered exercise hits the same topic (although not in exactly the same way) as the previous odd exercise. But the even-numbered answers are not provided. If a student is stuck on an even exercise, looking at the previous odd one (and its answer) can often provide the help needed.

As stated previously, more than 700 of our exercises include datasets  and randomization methods  available electronically. To ensure every student is able to read, analyze, interpret, and communicate data findings, we also provide ample exercises with full computer output.

We place all the exercises—including section-level exercises—at the end of the chapter. Our writing style is colloquial and encourages reading. We are telling a story about how to understand the world when you have data. Interrupting that story with exercises every few pages would encourage a focus on the calculations rather than the concepts.

Part Reviews. The book is partitioned into five conceptual parts; each ends with a Part Review. The part review discusses the concepts in that part of the text, tying them together and summarizing the story thus far. Then there are more exercises. These exercises have the advantage (for study purposes) of not being tied to a chapter, so they lack the hints of what to do that would come from that identification. That makes them more like potential exam questions and a good tool for review. Unlike, the chapter exercises, these are not paired.

Parts I-V Cumulative Review Exercises. A final book-level review section appears after the Part Review V. Cumulative Review exercises are longer and cover concepts from the book as a whole.

Additional Resources Online

Most of the supporting materials can be found online:

At the Pearson Math & Stats Resource Site

Within the MyLab Statistics course at [pearson.com/mylab/statistics](https://www.pearson.com/mylab/statistics)

Datasets are also available at dasl.datadesk.com.

Simulation and bootstrap applications (along with a few others) are available at MyLab Statistics and at astools.datadescription.com

Data desk RP is a statistics program with a graphical interface that is easy to learn and use. A student version is available at datadesk.com. Click on the **Teachers & Students** tab at the top of the page. Students beginning with the R statistics language may find it helpful to use Data desk’s ability to write out R code for plots and analyses such as those used in the text, thereby providing a graphical interface that can be easier for beginners. Students accessing datasets at DASL will find a quick link to Data desk.

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 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 MyLab Statistics kit, and StatCrunch is available by itself to qualified adopters. StatCrunch Mobile is also now available when you visit www.statcrunch.com from the browser on your smartphone or tablet. For more information, visit www.StatCrunch.com or contact your Pearson representative.

Additional Resources

Minitab® and Minitab Express™ make learning statistics easy and provide students with a skill-set that's in demand in today's data driven workforce. Bundling Minitab® software with educational materials ensures students have access to the software they need in the classroom, around campus, and at home. And having the latest version of Minitab ensures that students can use the software for the duration of their course. ISBN 13: 978-0-13-445640-9 ISBN 10: 0-13-445640-8 (Access Card only; not sold as standalone.)

JMP Student Edition is an easy-to-use, streamlined version of JMP desktop statistical discovery software from SAS Institute, Inc. and is available for bundling with the text. ISBN-13: 978-0-13-467979-2; ISBN-10: 0-13-467979-2



Pearson

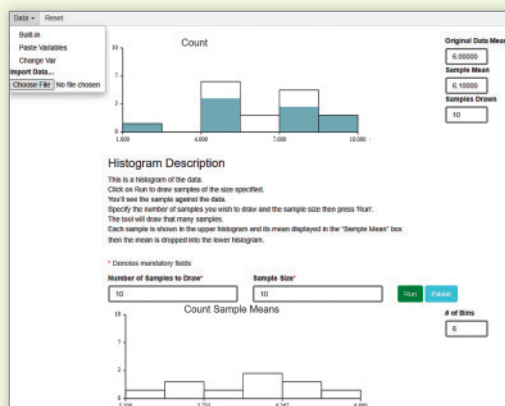
Resources for Success

MyLab[®] Statistics Online Course for *Intro Stats, 6e*
by Richard D. De Veaux, Paul F. Velleman,
and David E. Bock (access code required)

MyLab Statistics is available to accompany Pearson's market-leading text options, including Intro Statistics, 6e by De Veaux/Velleman/Bock (access code required). MyLab[™] is the teaching and learning platform that empowers you to reach every student. MyLab Statistics combines trusted author content—including full eText and assessment with immediate feedback—with digital tools and a flexible platform to personalize the learning experience and improve results for each student. Integrated with StatCrunch[®], an web-based statistical software program, students learn the skills they need to interact with data in the real world.

New exercises that incorporate REAL DATA.

MyLab Statistics exercises have been updated to include real data so students can understand the real-world implications of data analysis. Homework reinforces and supports students' understanding of key statistics topics within a real world context.



Exercises reflect diverse and relevant data and applications.

Examples and exercises throughout the textbook and MyLab Statistics use diverse and relevant data to help students understand how statistics applies to inclusive everyday life.

The screenshot shows the 'Assignment Manager' interface. It has tabs for 'Start', 'Select Media and Questions', and 'Choose Settings'. The 'Select Media and Questions' tab is active. It shows a list of 'Available Questions (6)' with checkboxes for selection. The 'Question Source' section on the right has checkboxes for 'Show publisher questions', 'Show Excel Project questions', 'Show additional test bank questions', 'Show custom questions (x) for this book', and 'Show other custom questions. Refine Selection...'. There are also buttons for 'Questions' and 'Media'.

Enhanced applications aid visualization and statistical understanding.

Applications have been updated and integrated into the Random Matters features and end of section problems. Students utilize applications to demonstrate randomness and illustrate sampling distributions through randomization techniques like bootstrapping. App problems have been created in the MyLab Statistics homework so students can think statistically about the output and communicate their understanding.

Minimum_Wage_Workers.txt				
StatCrunch • Applets • Edit • Data • Stat • Graph • Help •				
Row	Age	Workers	Group	var4
1	16-24	7978	Hourly Workers - Men	
2	25-34	9029	Hourly Workers - Men	
3	35-44	7696	Hourly Workers - Men	
4	45-54	7365	Hourly Workers - Men	
5	55-64	4092	Hourly Workers - Men	
6	65 and older	1174	Hourly Workers - Men	
7	16-24	7701	Hourly Workers - Women	
8	25-34	7864	Hourly Workers - Women	
9	35-44	7783	Hourly Workers - Women	
10	45-54	8260	Hourly Workers - Women	
11	55-64	4895	Hourly Workers - Women	
12	65 and older	1469	Hourly Workers - Women	
13	16-24	384	At or Below Minimum Wage - Men	
14	25-34	150	At or Below Minimum Wage - Men	
15	35-44	71	At or Below Minimum Wage - Men	
16	45-54	68	At or Below Minimum Wage - Men	
17	55-64	35	At or Below Minimum Wage - Men	
18	65 and older	22	At or Below Minimum Wage - Men	
19	16-24	738	At or Below Minimum Wage - Women	



Resources for Success

Student Resources

Intro Stats, 6th edition is part of De Veaux, Velleman, and Bock's Statistics series (ISBN-13: 978-0-13-680686-8; ISBN-10: 0-13-680686-4) This print textbook is available for students to rent for their classes.

Student's Solutions Manual provides detailed, worked-out solutions to odd-numbered exercises. This manual is available within MyLab Statistics.

Instructor Resources

Instructor's Solutions Manual (Download Only), contains solutions to all the exercises. These files are available to qualified instructors through Pearson Education's online catalog at www.pearson.com or within MyLab Statistics.

Online Test Bank and Resource Guide (Download Only), includes chapter-by-chapter comments on the major concepts, tips on presenting topics, extra teaching examples, a list of resources, chapter quizzes, part-level tests, and suggestions for projects. These files are available to qualified instructors through Pearson Education's online catalog at www.pearson.com or within MyLab Statistics.

TestGen® Computerized Test Bank (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 test bank are available for download from Pearson Education's online catalog at www.pearson.com.

PowerPoint® Lecture Slides: Free to qualified adopters, this classroom lecture presentation software is geared specifically to the sequence and philosophy of the book. Key graphics from the book are included to help bring the statistical concepts alive in the classroom. These files are available to qualified instructors through Pearson Education's online catalog at www.pearson.com or within MyLab Statistics.

Learning Catalytics: Learning Catalytics is a web-based engagement and assessment tool. As a "bring-your-own-device" direct response system, Learning Catalytics offers a diverse library of dynamic question types that allow students to interact with and think critically about statistical concepts. As a real-time resource, instructors can take advantage of critical teaching moments in the classroom or through assignable and gradeable homework.

ACKNOWLEDGMENTS

Many people have contributed to this book throughout all of its editions. This edition never would have seen the light of day without the assistance of the incredible team at Pearson. Director, Product Management Deirdre Lynch was central to the genesis, development, and realization of this project from day one. Our Content Manager, Suzanna Smith, has been invaluable in his support of this edition. Rachel Reeve, Content Producer, kept the cogs from getting into the wheels, where they often wanted to wander. Product Marketing Manager Alicia Wilson and Field Marketing Manager Demetrius Hall made sure the word got out. Media Producer Nicholas Sweeny put together a top-notch media package for this book. Senior Project Manager Rose Kernan led us expertly through every stage of production. Manufacturing Buyer Carol Melville, LSC Communications, worked miracles to get this book in your hands.

We'd also like to thank our accuracy checkers, Joan Saniuk, Stanley Seltzer, and Dirk Tempelaar, whose monumental task was to make sure we said what we thought we were saying.

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

Ann Cannon <i>Cornell College</i>	Sheldon Lee <i>Viterbo University</i>	Dirk Tempelaar <i>Maastricht University</i>
Susan Chimiak <i>University of Maryland</i>	Pam Omer <i>Western New England University</i>	Carol Weideman <i>St. Petersburg College</i>
Lynda Hollingsworth <i>Northwest Missouri State University</i>	Sarah Quesen <i>West Virginia University</i>	Ming Wang <i>University of Kansas</i>
Jeff Kollath <i>Oregon State University</i>	Karin Reinhold <i>SUNY Albany</i>	Lisa Wellinghoff <i>Wright State</i>
Cindy Leary <i>University of Montana</i>	Laura Shick <i>Clemson University</i>	Cathy Zucco-Teveloff <i>Rider University</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>	Robert L. Carson <i>Hagerstown Community College</i>	Jonathan Graham <i>University of Montana</i>
Froozan Pourboghnaf Afiat <i>Community College of Southern Nevada</i>	Jerry Chen <i>Suffolk County Community College</i>	Nancy Heckman <i>University of British Columbia</i>
Mehdi Afiat <i>Community College of Southern Nevada</i>	Rick Denman <i>Southwestern University</i>	James Helreich <i>Marist College</i>
Nazanin Azarnia <i>Santa Fe Community College</i>	Jeffrey Eldridge <i>Edmonds Community College</i>	Susan Herring <i>Sonoma State University</i>
Sanjib Basu <i>Northern Illinois University</i>	Karen Estes <i>St. Petersburg Junior College</i>	Mary R. Hudachek-Buswell <i>Clayton State University</i>
Carl D. Bodenschatz <i>University of Pittsburgh</i>	Richard Friary Kim (Robinson) Gilbert <i>Clayton College & State University</i>	Patricia Humphrey <i>Georgia Southern University</i>
Steven Bogart <i>Shoreline Community College</i>	Ken Grace <i>Anoka-Ramsey Community College</i>	Becky Hurley <i>Rockingham Community College</i>
Ann Cannon <i>Cornell College</i>		Debra Ingram <i>Arkansas State University</i>
		Joseph Kupresanin <i>Cecil College</i>

Kelly Jackson
Camden County College

Martin Jones
College of Charleston

Rebecka Jornsten
Rutgers University

Michael Kinter
Cuesta College

Kathleen Kone
*Community College of
Allegheny County*

Michael Lichter
*State University of
New York–Buffalo*

Susan Loch
University of Minnesota

Pamela Lockwood
*Western Texas A & M
University*

Wei-Yin Loh
University of Wisconsin–Madison

Steve Marsden
Glendale College

Catherine Matos
*Clayton College & State
University*

Elaine McDonald
Sonoma State University

Jackie Miller
The Ohio State University

Hari Mukerjee
Wichita State University

Helen Noble
San Diego State University

Monica Oabos
Santa Barbara City College

Linda Obeid
Reedley College

Charles C. Okeke
*Community College of
Southern Nevada*

Pamela Omer
Western New England College

Mavis Pararai
*Indiana University of
Pennsylvania*

Gina Reed
Gainesville College

Juana Sanchez
UCLA

Gerald Schoultz
Grand Valley State University

Jim Smart
*Tallahassee Community
College*

Chamont Wang
The College of New Jersey

Edward Welsh
Westfield State College

Heydar Zahedani
*California State University,
San Marcos*

Cathy Zucco-Teveloff
Rider University

Dottie Walton
*Cuyahoga Community
College*

Jay Xu
Williams College

INDEX OF APPLICATIONS

BE = Boxed Example; EM = Ethics Matters; E = Exercise; IE = In-Text Example; JC = Just Checking; RM = Random Matters; SBS = Step-by-Step Examples; SA = Student Activity; WCGW = What Could Go Wrong

Accounting

Troubleshooting, (E): 359

Advertising

Appliance Sales, (E): 317
Cell Phones, (JC): 443
Direct Mail, (E): 461
Endorsements, (E): 568
Political Ads, (E): 529, 530
Radio Advertising, (E): 560
Sexual Images in Advertising, (E): 634, 638–639
Super Bowl Commercials, (E): 380
Television Advertising, (E): 612

Agriculture

Chicken Feed, (E): 633–634
Egg Production, (E): 160, 633–634
Insect Control, (E): 383, (JC): 655
Livestock, (E): 489, 566, 640
Potatoes, (E): 459
Seed Viability, (E): 531
Vineyards, (E): 122, 324, 389

Banking

Credit Cards, (BE): 451–452, (E): 165, 276, 457, 458, 461, 491, 667, (JC): 520
Customer Age, (E): 667
Loans, (E): 558
Online Banking, (E): 423

Business (General)

Assets and Sales, (E): 333
CEO Compensation, (BE): 136–137, (E): 158–159, 491, 494, (IE): 108–109, 466–467, (WCGW): 149–150
Employee Injuries, (E): 459
Profits, (E): 332, 379
Women Executives, (E): 531
Women-Owned Firms, (E): 566, 568

Company Names

Allstate Insurance Company, (E): 495
Amazon, (BE): 4, (EM): 348, (E): 14, 429, (IE): 2, 3, 4, 6, 7
AT&T, (BE): 6
Bentley, (IE): 261
Cleveland Casting Plant, (E): 15

Daimler AG, (IE): 261
Facebook, (BE): 7, (E): 14, 422–423, 533, 534, (IE): 2, 402–403, 412, 445–446, (JC): 512, (SBS): 412–413
Ferrari World, (BE): 108
Ford Motor Company, (E): 15
Fukushima Daiichi Nuclear Power Plant, (BE): 25–26
GfK Roper, (E): 607
GlaxoSmithKline (GSK), (BE): 550, 553
Google, (BE): 7
Guinness Company, (IE): 468–469
Lowe's, (IE): 256
Mayo Clinic, (IE): 431
Nabisco Company, (E): 496
OkCupid, (IE): 67–68, 81
Preusser Group, (IE): 536
Rolls-Royce, (IE): 261
SmartWool, (BE): 501, (IE): 502–503
Snapchat, (EM): 660–661
Summit Projects, (BE): 501
Twitter, (IE): 402–403, 412, (SBS): 412–413
U.S. Small Business Administration Office of Advocacy, (EM): 689
White Star Line, (IE): 35

Consumers

Assembly Time, (E): 171
Cost of Living Index, (BE): 30
Credit Card Expenditures, (BE): 31, 38, 40
Pet Ownership, (E): 422
Purchases, (IE): 580–581, (SBS): 587–589
Shoes, (E): 154, 490
Tipping, (E): 329, 379, 387, 488, 489
Wardrobe, (E): 422

Demographics

Age of Couples, (IE): 624, 627
Deaths, (E): 120
Egyptian Body Structure, (E): 610
Emigration from State, (E): 459
Foreign-Born Citizens, (E): 603
Gender Status, (EM): 4, 271
Marital Status/Age, (BE): 106, 253, (E): 55, 120, 278, 282, 283, 284, 458, 529, 566, 711, 712
Population, (E): 63, 65, 119, (IE): 186–187, (JC): 265, 473–474, 478
Poverty, (E): 90

Race/Ethnicity, (EM): 271, (E): 90, 158, 529
U.S. Census, (E): 529

Distribution and Operations Management

Delivery Services and Times, (E): 96
Shipments, (E): 154

E-Commerce

Online Shopping, (E): 14, 198, 238–239
Profits, (E): 379

Economics

Childhood Household Financial Situation, (SA): 674
Cost of Living, (BE): 175–176, 182, 185, (E): 118, 125, 242
Crowdedness, (E): 286, 287
Financial Situations of Most Americans, (E): 54
GDP, (E): 287–288
Gross Domestic Product, (E): 200–201
Human Development Index (HDI), (E): 279
Income, (JC): 39
Income and Housing Costs, (E): 201
Inflation, (E): 285
Labor Force Participation Rate, (E): 634
Market Segments, (E): 276
Occupy Wall Street Movement, (E): 569
Prices, (JC): 265
United Nations Development Programme (UNDP), (E): 279

Education

Academic Performance, (IE): 361–362
Admissions/Placement, (BE): 83–84, (E): 96–97, 154, 165–166, 167–168, 424
Attainment by Age Group, (E): 673
Cheating, (E): 567
Childhood Household Financial Situation, (SA): 674
College and Financial Well-Being, (E): 91
College Attendance Rates, (E): 529
College Retention rates, (E): 462
College Value, (E): 87, 88, 89, 96
Core Plus Mathematics Project (CPMP), (E): 608–609
Cornell University, (IE): 373

Cost of Higher Education, (E): 638, 718
 Cramming, (E): 165, 326
 Dartmouth College, (E): 490
 Earnings of College Graduates, (E): 707, 708–709
 Educational Testing Service (ETS), (JC): 133
 Employment of College Students, (IE): 520
 Grade Levels, (E): 14
 Grades/Scores/GPA, (BE): 693, (E): 61, 64, 65, 121, 122, 123, 153, 154, 155, 156, 157, 159–160, 239, 277, 281, 317, 318–319, 320, 323, 326, 331, 358, 379, 382, 383, 458, 488, 494, 528, 529, 565, 568, 570, 672, 717, (IE): 142, 470, (JC): 131, 133, 161, 182–183, 206, (SBS): 134, 138–139, 143, 144
 Graduation Rates, (E): 121, 427–428, 459, 462
 Group Projects, (E): 424
 Height and Reading, (E): 201
 High-School Dropout Rate, (E): 532
 High-School Graduation Rate, (E): 605
 Homecoming, (E): 388
 Kindergarten, (E): 160
 Living arrangements, (JC): 662
 LSAT, (E): 488
 Magnet Schools, 94, (E): 94
 Major Field, (E): 169
 Meal Plans, (E): 490
 Mortality and Education, (E): 717
 National Center for Education Statistics, (E): 531
 Parental education, (E): 531
 Post-Graduation Activities, (E): 92–93, (IE): 650–652, 654, (SBS): 652–653
 Prerequisites, (E): 427
 Professors, (E): 157
 Public Opinion, (E): 164
 Reading, (E): 281, 383, 386, 610, (IE): 557
 School Absenteeism, (E): 531
 School Mail, (E): 169
 School Records, (E): 15
 School Uniforms, (E): 459
 Scorecard Analysis, (SA): 66, 97, 127, 162
 Software, (E): 386, 559–560
 Spring Break, (E): 458
 Student Evaluations, (IE): 373–374
 Student Goals, (IE): 407–409
 Studying, (E): 169
 Summer School, (E): 611, 637
 Texas A&M University, (E): 356
 University of California at Berkeley, (BE): 83–84, (E): 96–97
 University of Texas, (E): 332
 Williams College, (E): 331–332, (IE): 98, 107

Energy

Batteries, (E): 426, 570
 Energy Information Administration (EIA), (IE): 252

Fuel Economy, (E): 120–121, 123, 154, 201, 202, 234, 244–245, 286, 287, 289, 359, 383, 389, 633, 637–638, 713, 714, (IE): 148–149, 261–262, (JC): 39, (SBS): 40–41
 Fukushima Daiichi Nuclear Power Plant, (BE): 25–26
 Gas Prices, (E): 63, 120
 Oil Prices, (IE): 251–252
 Wind Power, (E): 497

Environment

Acid Rain, (E): 64, 164–165, 532
 Camping, (E): 119
 City Climate, (E): 715
 Cloud Seeding, (E): 123, 124
 Earthquakes and Tsunamis, (BE): 25–26
 Emissions Testing, (E): 462, 559
 Environmental Protection Agency, (E): 154
 Floods, (E): 61
 Global Warming, (BE): 450, (E): 14, 60, 242–243, 386, 713–714
 Hard Water, (E): 611, 614
 Hurricanes, (BE): 210, 219, 227, (E): 63, 200, 280, 669, 709, (IE): 107, 173, (JC): 303, 322
 National Hurricane Center (NHC), (IE): 173, 177
 National Oceanic and Atmospheric Administration (NOAA), (IE): 173
 National Weather Service, (IE): 107
 Northeast Regional Climate Center, (E): 495
 Old Faithful, (E): 167, 328, (SA): 498
 Ozone, (E): 121, 715
 Pollution, (E): 531
 Rainfall, (E): 495, 634
 Sea Ice, (E): 709–710
 Seasons, (E): 166–167
 Snow, (E): 490
 Soil, (E): 357
 Streams, (E): 16, 165, 200, 325, 326, 569, 610, 712, 713
 Temperature, (E): 153, 155, 200, 281, 330, 636, 715
 Tornadoes, (E): 61
 Typhoons, (IE): 173
 Water, (E): 166, 235, 236, 244
 Weather, (BE): 210, 219, 227, (E): 14, 63, 94, 116, 153, 172, 198, 200, 280, 423, 490, 495, 669, 709, (IE): 107, 173–175, (JC): 303
 Wildfires, (E): 170–171, 239–240
 Wind Speed, (E): 156, 286, 316–317, 635, 636, (IE): 98–102, 106, 107, (JC): 322

Bacon, Francis, (IE): 256, 682
 Bayes, Thomas, (BE): 418, (IE): 417
 Bernal, Egan, (JC): 8
 Bernoulli, Jacob, (IE): 393
 Berra, Yogi, (IE): 174, 399
 Bohr, Neils, (IE): 250
 Box, George, (IE): 137, 208
 Boyle, Robert, (E): 288
 Brahe, Tycho, (IE): 9
 Buchanan, Pat, (IE): 254–255
 Bush, George W., (IE): 254, 256, 349
 Carroll, Lewis, (IE): 1, 413
 Ceci, Stephen, (IE): 373–374
 Clinton, Bill, (E): 568
 Clinton, Hilary, (E): 565
 Collier, Wayne, (E): 389
 Cornet, Henri, (JC): 8
 de Moivre, Abraham, (IE): 135, 434, 435
 D'Ignazio, Catherine, (EM): 484
 Descartes, René, (BE): 176
 Efron, Bradley, (BE): 479
 Einstein, Albert, (E): 492
 Fechner, Gustav, (IE): 363
 Fisher, Ronald Aylmer, (BE): 187, 543, 549, (IE): 469*n*, 472, 503
 Fleet, Frederick, (IE): 18
 Franklin, Benjamin, (IE): 396
 Froome, Christopher, (JC): 8
 Gallup, George, (IE): 337
 Galton, Francis, (BE): 217, (EM): 217–218
 Garin, Maurice, (JC): 8
 Gauss, Carl Friedrich, (BE): 209, (IE): 135, 469*n*
 Gill, Colin, (E): 388
 Gore, Al, (IE): 254–255
 Gosset, W. S., (IE): 468–469, 470
 Grant, U.S., (IE): 396
 Gretzky, Wayne, (E): 63
 Halifax, Lord, (IE): 362
 Hamilton, Alexander, (IE): 396
 Harvey, William, (IE): 255
 Homer, (E): 667–668
 Howe, Gordie, (E): 63
 Hume, David, (IE): 541
 Jackson, Andrew, (IE): 396
 Jastrow, J., (IE): 364
 Jefferson, Thomas, (IE): 396
 Johnson, Gary, (E): 565
 Johnson-Thompson, Katarina, (IE): 128, 129, 130
 Kepler, Johannes, (IE): 9, 10
 Klassen, Cindy, (SBS): 621
 Klein, Lauren, (EM): 484
 Kohavi, Ronny, (BE): 4, (IE): 2
 Landers, Ann, (BE): 350
 Landon, Alf, (IE): 336, 337, 352
 Laplace, Pierre-Simon, (IE): 465
 Legendre, Adrien-Marie, (BE): 209
 Lincoln, Abraham, (IE): 396
 Lowell, James Russell, (IE): 507
 Maas, Jim, (IE): 512

Famous People

Archimedes, (IE): 255
 Armstrong, Lance, (JC): 8

McGwire, Mark, (E): 169
 Meir, Jessica, (IE): 248
 Michelson, Albert Abraham, (E): 492
 Moore, David, (IE): 366*n*
 Munchausen, Baron, (IE): 479*n*
 Nader, Ralph, (IE): 254–255
 Neubauer, Peter, (EM): 628
 Newton, Isaac, (BE): 176, (IE): 10
 Nibali, Vincenzo, (JC): 8
 Nobel Laureates, (E): 14
 Obama, Barack, (E): 387, 558, 568
 O’Loughlin, William Francis Norman, (IE): 32
 O’Neil, Cathy, (EM): 484
 Pierce, C. S., (IE): 364, 369*n*
 Poganis, Paul, (IE): 248
 Quenouille, Maurice, (BE): 479
 Raspe, Rudolf Erich, (IE): 479*n*
 Robbins, Rebeca, (IE): 512
 Rodriguez, Alex, (E): 63
 Roosevelt, Franklin Delano, (IE): 336, 337, 352
 Rudder, Christian, (IE): 67
 Ruth, Babe, (E): 169
 Saunderson, Nicholas, (IE): 418
 Sophocles, (IE): 323
 Spicer, Sean, (BE): 70
 Stein, Jill, (E): 565
 Stigler, Steven, (BE): 418
 Thiam, Nafissatou, (IE): 128–130, 150
 Thomas, Geraint, (JC): 8
 Trousseller, Louis, (JC): 8
 Trump, Donald J., (E): 565
 Truzzi, Marchello, (IE): 520
 Tufte, Edward, (WCGW): 45*n*
 Tukey, John W., (BE): 27, 102, 479, (IE): 264–265, 443
 Van Buren, Abigail, (BE): 394
 Venn, John, (IE): 397
 Wanamaker, John, (IE): 499
 Washington, George, (IE): 396
 Watson, William Albert, (IE): 32
 Wayne, John, (E): 532
 Weeks, David, (E): 388
 Wiggins, Bradley, (JC): 8
 Wunderlich, Carl, (IE): 514

Finance and Investments

Assets, (E): 124
 Brokerage Accounts, (E): 668
 Computer Lab Fees, (E): 532
 Currency, (BE): 409, (E): 558, (IE): 395–396
 Drug Development Costs, (E): 529
 401(k) Plans, (E): 15
 Interest Rates, (EM): 304, (E): 201, 238
 Mutual Funds, (E): 233
 Profits, (E): 171
 Saving and Investment Performance, (BE): 100
 Second Jobs, (E): 558

Stocks, (E): 325, 565
 T Bill Rates, (E): 282, 283, 287
 Websites, (E): 528

Food/Drink

Alcohol Consumption, (E): 59, 331, 380, 382–383, 388–389, 428, 567, (IE): 414–415
 Appetite, (E): 612
 Bananas, (E): 163
 Beer, (IE): 468–469
 Bread, (E): 164, 569
 Brewpubs, (E): 57
 Burgers, (E): 202
 Caffeine, (E): 122
 Candy, (E): 318, 321, 425, 530, 668, (SBS): 405–407, (SA): 719–720
 Cereal, (E): 60, 118, 172, 234, 320, 609, 714–715, (IE): 226–227, 249–250, (SBS): 145–147, 224–226
 Chicken, (E): 427
 Coffee, (E): 199
 Cookies, (E): 388, 496, 533, 561
 Cooking, (E): 422
 Cranberry Juice, (E): 671
 Diet, (E): 607
 Fast Food, (E): 607
 Fish, (BE): 471, 473, 511, (E): 488, 492, 561, 672
 Food Preferences, (E): 607
 Grocery Shopping, (E): 14
 Hams, (E): 155
 Hot Dogs, (E): 532, 607, 710, 711
 Meals, (E): 331, 569, (JC): 662
 Milk, (E): 357
 Nutritional Data, (BE): 211–212, (E): 241, 278, 318, 492–493, 529, 532, 607, 608, 710, 711, (IE): 207–208, 209–210, 219–220, 221, 223, 308–309, (WCGW): 227–228
 Nuts, (E): 668
 Pet Food, (BE): 365, 370, (E): 381, 385
 Pizza, (E): 61, 62, 117, 532
 Quality Control, (BE): 365, 370, (E): 359
 Restaurants, (E): 428
 Safety, (BE): 471, 473, 511, (E): 386, 461
 Salt, (E): 172
 Seafood, (E): 460
 Snack Foods, (E): 357, 495–496, 533, 569
 Soft Drinks, (E): 529, (IE): 557
 Soup, (BE): 581–582, 583–584, (EM): 584, (E): 458, (IE): 336–338
 Thirst, (E): 612
 Tomatoes, (E): 160, 379, 381
 Unsafe Food, (E): 427
 Yogurt, (E): 496, 533, 637

Games

Cards, (E): 166, 425
 Coin Flips, (E): 530, (IE): 519–520, (SA): 430

Coin Spins, (E): 460, 560
 Coin Toss, (BE): 393–394, (E): 422
 Dice, (E): 422, 530, 668
 Gambling, (BE): 394, (E): 423, 461, 528, (IE): 694
 Keno, (BE): 394
 Lottery, (E): 425, 495, 669–670, (JC): 429
 Roulette, (E): 423
 Scrabble, (E): 565
 Spinners, (E): 423

Government, Labor, and Law

Arrests, (E): 672, 673
 Bureau of Labor Statistics, (BE): 26
 Car Thefts, (E): 199
 Checkpoints, (E): 459
 Corruption, (E): 205
 Crime and Television Watching, (E): 199
 Death Penalty, (E): 461, (SBS): 448–449
 False Conviction, (IE): 548
 Fraud Detection, (E): 165
 Freedom, (E): 205
 Identifiers in Data, (EM): 7
 Juries, (E): 531, (IE): 501–502, 503
 Labor Force Participation Rate, (E): 634
 National Security Agency (NSA), (IE): 7*n*
 NYPD, (E): 671
 Parks, (E): 358
 Parole, (E): 460
 Paycheck Protection Program (PPP), (BE): 585–586
 Polygraphs, (E): 428
 Prisons, (E): 116, 461
 Race and Police Action, (E): 90
 Radon Testing, (E): 385
 Roadblocks, (E): 357
 Speeding, (E): 424, 568
 Taxes, (E): 424
 Traffic Stops, (BE): 656, 658–659, (E): 90
 U.S. Census Bureau, (BE): 339, (JC): 39, 473–474, 476, 497
 U.S. Postal Service, (E): 14
 Violence against Women, (E): 669
 Zip Codes, (E): 14, 64

Human Resource Management/Personnel

Absenteeism, (E): 428
 Career Success, (E): 91
 College Student Employment, (IE): 520
 Commute Times, (E): 609, (JC): 142, (RM): 35–36, 43, 139–141
 Flexible Work Schedules, (BE): 618, 620–621, 623–624
 Hiring, (E): 459, 462
 Hours Worked, (BE): 26
 Human Resource Data, (E): 564

Job Discrimination, (E): 559
 Job Growth, (E): 65
 Job Hunting, (E): 14
 Job Satisfaction, (E): 92, 356, 357, 359, 603, 611, 637
 On-Site Day Care, (E): 528
 Placement Scores, (IE): 557
 Rating Employees, (E): 233
 Sick Days, (E): 61
 Workplace Ratings, (E): 97

Insurance

Auto Insurance, (E): 423, 495, 635, 636
 Health Insurance, (E): 168, 427, 568–569

Manufacturing

Assembly Time, (E): 171
 Beer, (IE): 468–469
 Cars, (E): 460, 668, 671
 Chips, (E): 333
 Computers, (E): 123
 Electronics, (E): 428
 Emergency Shutoff, (E): 564
 Machine Settings, (BE): 596
 Metals, (E): 15, (IE): 499–502, 503–504
 Pottery, (E): 199
 Rivets, (E): 157
 Safety, (E): 383
 Shoes, (E): 57, 381–382, (IE): 368
 Skates, (JC): 133
 Swimsuits, (E): 382

Marketing

Direct Marketing, (E): 457
 Packaging, (E): 534, 570
 Social Media, (EM): 660–661

Media and Entertainment

American Journal of Health Behavior, (IE): 414
Archives of General Psychiatry, (E): 606
 Belmont Report, (EM): 4, 68
Berkshire Eagle, (IE): 107
British Medical Journal, (EM): 188, (E): 671
Chance magazine, (E): 673
 Concerts, (E): 276, 277
Consumer Reports, (BE): 5, 8, (E): 427, 607, 710
Data Feminism, (EM): 484
The Economist, (IE): 6
 “Ethical Principles and Guidelines for the Protection of Human Subjects of Research”, (EM): 4
Journal of Applied Psychology, (E): 611–612
Journal of the American Medical Association, (E): 606

The Lancet, (E): 605
Literary Digest, (IE): 337, 352
 Motion Picture Association of America (MPAA), (E): 54
 Movies, (E): 54–55, 58, 59, 62, 87–88, 89, 90, 93, 172, 279, 315–316
 Music, (E): 124, 155, 164, 169, 276, 277, 382, 422, 614, 634, (IE): 519
New England Journal of Medicine, (E): 670
The New York Times, (BE): 349, (IE): 107
 News Reporting, (BE): 445, 446–447, (EM): 188, (E): 14, 58, 90, 117, 565, 604
 Online Magazines, (E): 531
 Public Opinion Surveys, (IE): 336
Readers' Digest, (E): 388
 Rock Concerts, (E): 118, 157
 Social Networking, (BE): 7, (EM): 660–661, (E): 14, 357, 422–423, 533, 534, 604, (IE): 2, 412, 445–446, (JC): 512, (SBS): 412–413
Sports Illustrated, (E): 236
 Televisions, (E): 329
Three Identical Strangers, (EM): 628
The Twinning Reaction, (EM): 628
Weapons of Math Destruction, (EM): 484
Wired, (IE): 2
 World Happiness Report, (E): 160, 236, 329
World Drug Report, (E): 709

Pharmaceuticals, Medicine, and Health

Adolescent Dangerous Behavior, (IE): 339, (WCGW): 44
 Alcohol Consumption, (E): 382–383, 388–389, (IE): 414–415
 Alternative Medical Treatments, (BE): 371, (E): 357, 358, 381, 389
 Alternative Medicine, (E): 15, 169
American Journal of Health Behavior, (IE): 414
 Antacids, (E): 388
 Aspirin, (E): 528, (JC): 512
 Baldness, (E): 201
 Blindness, (E): 15
 Blood Pressure, (E): 94, 380, 382, 426, 427, (SBS): 183–184
 Blood Types, (E): 89, 425, 605
 Body Fat, (BE): 136, 686–687, 695–696, (E): 15, 56, 57, 62, 243, 320, 493, 494, 534, (IE): 281, 292–297, 675–676, 677, 679, 680, 687, 688–689, 694, (RM): 440, (SBS): 682–683, (WCGW): 302
 Body Mass Index (BMI), (BE): 437–438, (E): 534, (IE): 696–699
 Body Temperature, (E): 160, 491, 492, 532, (RM): 515–516, 538
 Brain, (E): 637, 714
 Brain Supplement, (EM): 521–522
 Caffeine, (E): 122
 Cancer, (BE): 187, (E): 88–89, 380, 386, 532, 569, 605, (IE): 187, (SBS): 74–76
 Caring for Household Members, (E): 125, 126, 161, 493
 Carpal Tunnel Syndrome (CTS), (E): 605, (JC): 586, 589
 Causes of Death, 94, 90, (E): 90
 Centers for Disease Control and Prevention, (BE): 437, (E): 59, 279, 605, (IE): 131, (WCGW): 44
 Cholesterol, (E): 122, 123, 159, 426, 427, 490, 491, 534, 610, 614, 668, 711
 Color-Blindness, (E): 565
 Congenital Abnormalities, (E): 530–531
 COVID-19, (BE): 573–574, 575, 580, 585–586, (E): 16, 89, 94, 429, 462, 605, (IE): 335–336
 Death Rates, (E): 383
 Depression, (E): 330
 Diabetes, (BE): 536–537, 541, 542, 549–550, 553
 Dialysis, (E): 164
 Diet, (E): 88–89, 89, 92, 95–96, 380, 382, 387
 Diseases/Illnesses/Injuries, (BE): 187, 536–537, 541, 542, 547, 549–550, 553, (E): 60, 88–89, 95, 119, 201, 277, 379, 380, 381, 382, 386, 388, 461, 462, 532, 559, 560, 569, 605, 606, 668, 670, (IE): 557, 655–656, 696–699, (JC): 586, 589, (SBS): 74–76, 657–658, 659–660
 Domoic Acid, (E): 357
 Drug Abuse, (E): 202, 240–241, 357, 379
 Drug Development, (E): 529
 Drug Use/Abuse, (E): 709
 Eating Disorders/Weight Issues, (E): 94–95, 387, 530, 568, 605, 606, 634, 639
 Emotional Health, (E): 330, 380, 383, 386, 606
 ESP, (E): 236
 Exercise, (E): 15, 94–95, 380, 381, 382
 Fertility, (E): 285
 Food and Drug Administration (FDA), (BE): 364, 537, 553
 Gastric Freezing, (JC): 368
 Gene Therapy, (E): 386
 Genetics, (E): 328
 Gestation/Pregnancy/Childbirth, (BE): 471, 511, 643, 646, 654, (EM): 418–419, (E): 15, 169, 232, 242, 283, 324, 331, 381, 385, 422, 426, 427, 495, 531, 564, 565, 567, 606, 666–667, 670, 673, (IE): 431–432, 440–441, 442–443, 464–465, 475–476, 509–510, (RM): 185–186, 341–343, 444, 481–482, (SBS): 474–475, 507–509, (WCGW): 452–453
 Harvard School of Public Health, (IE): 414
 Health Records, (E): 14
 Heart Disease, (BE): 541, 542, 547, 549–550, 553, (E): 60, 201, 379

Height and Weight, (BE): 178–179, 437–438, (E): 164, 200–201, 320, 379, (IE): 131–133, 148, 179–181, 216–217, 261–262, 302, (JC): 137

Hepatitis C, (IE): 655–656, (SBS): 657–658, 659–660

Hippocratic Oath, (EM): 484

HIV Testing, (E): 428

Hospitals, (E): 96, 119, 383

Insomnia, (E): 380, 381

Journal of the American Medical Association, (IE): 541

Life Expectancy, (E): 201, 278, 285, 289, 290, 334, (IE): 257–258

Mammograms, (EM): 554, (E): 606

Manual Dexterity, (IE): 617, 619, (SBS): 625–626, (SA): 641

Marijuana, (E): 64

Mayo Clinic, (IE): 431

Medical Testing, (E): 559

Medical Treatments, (JC): 371, 384

Medication Side Effects, (E): 529

Medication Trials, (BE): 536–537, 541, 542, 549–550, 553, (E): 14, 95, 200, 386, 388, 528, 530, 605, (IE): 557

Memory, (E): 610–611, 614

Menopause, (E): 380

National Center for Biotechnology, (BE): 363

National Center for Chronic Disease Prevention and Health Promotion, (IE): 339

National Center for Health Statistics (NCHS), (E): 14, (IE): 431, 464

National Health and Nutrition Examination Survey (NHANES), (BE): 131, (IE): 131

National Institutes of Health, (BE): 363, (IE): 696

New England Journal of Medicine (NEJM), (BE): 536, 541, 547, 549, (E): 670

Omega-3, (E): 381

Pain, (E): 605, 607

Pulse Rates, (E): 609

Sleep, (E): 62, 205, (IE): 512–513, 577–578, (RM): 76–77, (SBS): 512–513, 578–579

Sleep Foundation, (SBS): 512

Smoking, (E): 90–91, 123, 279–280, 331, 386, 388, 529, 565, 567, 606, (IE): 110–111, 187

Stress Testing, (E): 14

Tattoos, (E): 95, (IE): 655–656, 659–660, (SBS): 657–658

TB Screening Test, (SBS): 416–417

Therapeutic Touch (TT), (IE): 541–542

Treatment Modalities, (E): 60

University of Texas Southwestern Medical Center, (IE): 655

Vaccinations, (E): 605

Veterinary Medicine, (E): 558

Vision, (E): 565

Vitamins, (E): 380, 383, 386, 461

Women's Health Initiative, (BE): 363

Politics and Popular Culture

Approval Ratings, (E): 558

Arrests, (E): 672

Convention Bounce, (E): 607

Crime and Television Watching, (E): 199

Election Polls, (E): 565, 566, 604, 606, 607, (IE): 260–261, (SA): 463, 615–616

Embrace vs. Protect Attitudes, (E): 718

Liberalism/Conservatism, (E): 93, 94, 95, 389, 668

Petitions, (E): 568–569

Pets, (E): 422, 567, 605

Political Parties, (E): 88, 89, 427, 671–672

Roller Coasters, (BE): 102–103, 108, (E): 200, 203, 204, 235, 236, 358, (IE): 304–307, 692

Socks, (E): 564

Statue of Liberty, (E): 168

Television Violence, (E): 611–612

Titanic Sinking, (BE): 70–71, 82, (E): 172, 422, 671, 673, (IE): 18–20, 25, 28, 32–37, 39, 78–79, (RM): 23–24, (WCGW): 44, 46, 84

Voting and Elections, (E): 14, 357, 358, 389, 426, 565, 566, 604, 606, 607, (IE): 254–256, 336, 337, 352, (SA): 206, 246, 290–291, 322, 615–616

Zodiac Signs, (IE): 642, (SBS): 646–648

Quality Control

Airplanes, (IE): 520

Assembly Line, (E): 559

Bottling, (E): 386

Brewing, (IE): 468–469

Cars, (E): 564

Catheters, (E): 560

Chips, (E): 560

Fireworks, (E): 386

Food Inspection and Safety, (BE): 511, (E): 359

Light Bulbs, (E): 425

Oranges, (E): 386

Pet Food, (BE): 365, 370

Product Inspections and Testing, (E): 531, 564, 567, 570

Product Ratings and Evaluations, (E): 607, 608, (JC): 39

Recalls, (E): 570

Underground Storage Tanks, (E): 564

Real Estate

Home Ownership Rate, (E): 559

Housing Prices, (BE): , 294, (E): 201, 237–238, 315, 317, 318, 319, 333–334, 489, 494–495, 609, 710, 717–718, (JC): 218, 222, (SBS): 300–302

Mortgages, (EM): 304, (E): 201, 238

Racial Discrimination, (E): 673

Vacant Houses, (E): 529

Salary and Benefits

CEO Compensation, (BE): 136–137, (E): 158–159, 491, 494, (IE): 108–109, 466–467, (RM): 476–477, (WCGW): 149

Earnings of College Graduates, (E): 707, 708–709

Employee Benefits, (E): 427

Income and Housing Costs, (E): 201

Salaries/Payroll, (BE): 136–137, (E): 61, 93, 154, 155, 158–159, 170, 198, 237, 288, 319, 329, 489, 639, (IE): 108–109, (WCGW): 149

Sales and Retail

Appliances, (E): 317

Assets and Sales, (E): 333

Books, (E): 197, 232–233

Coffee, (E): 199

Customer Service, (E): 154

Diamonds, (BE): 309–311

Discounts, (E): 423–424

Groceries, (E): 14, 489, 558

Online Shopping, (E): 14, 198, 238–239

Profits, (E): 715, 716

Purchase Amounts, (E): 603–604

Sales, (E): 64

Sales Promotions, (E): 425

Shopping, (E): 422

Socks, (BE): 501, (IE): 502–503

Science

Abalones, (E): 277

Alligators, (E): 244

Arm Length, (E): 359

Asteroid Impact, (E): 90

Astronomy, (E): 117, 204, 288–289

Birds, (E): 60, 63–64, 235, 490, 557, (JC): 265

Cattle, (E): 154, 156, 157, 159

Cigarettes, (E): 236–237

Craters, (BE): 680–681, 687–688

Crocodiles, (E): 328

Deer Ticks, (E): 462

Dexterity Testing, (E): 198, 320–321, (IE): 263

Diet, (SBS): 74–75

Dowsing, (E): 383, 530

Draining Tanks, (E): 332

Eggs, (E): 160

Elephants, (E): 284

Eye Color, (E): 121

Fish, (E): 359

F/stops, (IE): 189, 191, (WCGW): 192

Genetics, (E): 669, (IE): 650, (JC): 73

Hair Color, (E): 121

Hamsters, (E): 565
 Handedness, (E): 424, (SBS): 438–439
 Hippos, (E): 284
 Identity vs. Privacy, (EM): 77–78
 Language, (E): 567
 Laundry Detergent, (E): 383, 387
 Lumber, (E): 289
 Manatees, (E): 325, 326
 Mazes, (E): 496, 533
 Mouth Volume, (JC): 689
 Nail Polish, (IE): 370, 372, (SBS): 366–367
 Oranges, (E): 290
 Paired Studies, (EM): 628
 Pendulums, (E): 288
 Penguins, (E): 334, (IE): 248, (SBS): 265–268
 Pi, (E): 669
 Pottery Glazes, (E): 386, 561
 Reaction Times, (SA): 384, 535
 Shoe Size and IQ, (IE): 256, (WCGW): 192–193
 Speed of Light, (E): 492
 Stigler, Stephen M., (E): 492
 Temperatures, (SBS): 103–104
 Time Judgments, (E): 388
 Trees, (E): 157, 235, 290, 328
 Vision, (E): 328
 Walking in Circles, (E): 16
 Weighing Bears, (E): 15
 Youthful Appearance, (E): 388

Service Industries and Social Issues

Adoption, (E): 57
 Athletics and Relationships, (BE): 401, 409, 410
 Cantril Scale, (E): 457
 Charitable Solicitations, (E): 457, 458, 461, 531, (JC): 577
 Crawling, (E): 491, 716–717
 Donors, (E): 158
 Families, (E): 324
 Fundraisers, (E): 570
 Gun Control, (E): 92
 Honesty, (BE): 590, (E): 15, 380
 Internet Surfing, (E): 330, (IE): 596
 Leisure Time, (E): 126, 161, 493
 Living Arrangements, (E): 566
 Marriage, (BE): 106, 253, (EM): 111, (E): 55, 120, 278, 282, 283, 284, 458, 529, 566, 711, 712
 Men's Attitudes, (E): 607
 Occupy Wall Street Movement, (E): 569
 Online Dating, (IE): 67–68, 81
 Paralyzed Veterans of America, (E): 461, 531
 Parenting, (E): 604, 607
 Psychics, (E): 528
 Pubs, (E): 388–389
 Racial Discrimination, (BE): 656, 658–659, 660, (E): 669, 673

Relaxing, (E): 493
 Religion, (E): 15, 567
 Safety at Play, (E): 567
 Sharing Personal Information, (E): 604
 Social Life, (E): 205, 357, 493
 Social Networking, (BE): 7, (EM): 660–661, (E): 14, 422–423, 533, 534, 604, (IE): 2, 412, 445–446, (JC): 512, (SBS): 412–413
 Violence against Women, (E): 669
 Working Parents, (E): 672
 World Happiness, (E): 160–161, 205

Sports

Age of Players, (E): 604
 Archery, (E): 564
 Athlete Name Recognition, (E): 568
 Baseball, (BE): 643, 646, 654, (E): 63, 123–124, 169, 202–203, 237, 288, 317–318, 357, 379, 389, 460, 611, 614, (SBS): 517–519
 Basketball, (E): 120, 422, 561
 Bowling, (E): 425
 Exercise, (E): 637
 Fans, (E): 558
 Fishing, (E): 566–567
 Football, (BE): 69–70, 72–73, (E): 14, 60–61, 158, 245, 331–332, 380, 422, 531
 Golf, (E): 493, 496–497, 529, 614
 Hockey, (E): 63, 707, 708, 709
 Indy 500, (E): 16, 17
 Injuries, (E): 60
 Kentucky Derby, (E): 16, 17, 122, 199, (IE): 27–28, 39, (WCGW): 46
 Motorcycle Riding, (BE): 258–260, 307–308, (IE): 536, (SBS): 539–540
 Olympics, (BE): 133, (E): 59, 118, 157, 168, 243–244, 330, (IE): 128–130, 626–627, (SBS): 621–622, (WCGW): 150
 Relationships, (BE): 401, 409
 Running, (E): 60, 317, 318, 321, 613, 636
 Skiing, (E): 118, 157, 496, 533
 Skydiving, (E): 384
 Soccer, (E): 163–164, 427, 566
 Speed Skating, (E): 168, (IE): 626–627, (SBS): 621–622
Sports Illustrated, (E): 236
 Super Bowl, (BE): 69–70, 72–73, (E): 14, 60–61, 380
 Swimming, (E): 284, 383, 612–613
 Tour de France, (E): 17, 285–286, (JC): 8
 Uniforms, (E): 427, 566
 Weightlifting, (E): 289
 Wheelchair Marathons, (E): 640

Surveys and Opinion Polls

American Association for Public Opinion Research (AAPOR), (EM): 447, (JC): 411

American Community Survey, (JC): 473
 American Time Use Survey (ATUS), (BE): 26, (E): 205
 Annenberg Inclusion Initiative, (EM): 87
 Approval Ratings, (E): 462
 Athlete Name Recognition, (E): 568
 Car Purchases, (E): 605
 Cell Phones, (E): 422
 Deloitte, (IE): 68
 Design, (SA): 360
 Elections, (E): 565, 566, 604, 606, 607, (SA): 615–616
 Gallup Polls, (BE): 21, 450, (E): 357, 427, 457, (IE): 337, (SBS): 448–449
 GfK Roper, (E): 425–426
 International Bedroom Poll, (IE): 577
 Liberalism/Conservatism, (E): 389
Literary Digest, (IE): 352
 Margin of Error, (E): 459
New York Times, (BE): 349
 Pew Research Organization, (BE): 21–22, 338, 445, 446–447, (E): 54, 55, 92, 381, 389, 424, 604, (JC): 405, 443, (SBS): 402–403
 Phone Surveys, (E): 359
 Postpurchase Surveys, (E): 171
 Presidential Popularity, (E): 568
 Public Opinion Polls, (E): 164, 358, 424, 425–426, 459, 462, 604, (IE): 335–336, (JC): 411
 Random Samples, (E): 358
 Student Surveys, (E): 66, 88, 93, 172, 356, 357, 490
 Zip Code, (E): 201

Technology

Cell Phones, (BE): 80–81, (E): 201, 277, 359, 387, 422, 459, 530, (JC): 443
 Chips, (E): 333
 Computers, (BE): 8, (E): 424, 567–568, 716
 Data Collection, (EM): 4–5
 Data Ownership, (EM): 42
 Device Usage, (BE): 21–22
 Disk Drives, (E): 197, 233, 234
 E-Mail, (EM): 554, (E): 56, 57, 558–559
 E-Readers, (E): 428
 Identity vs. Privacy, (EM): 77–78
 Internet, (BE): 9, (E): 14, 58, 64, 117, 330, 390, 457, 458, 460, 604, (IE): 596
 MP3 Players, (E): 156
 National Strategy for Trusted Identities in Cyberspace, (BE): 7
 Online Dating, (IE): 67–68, 81
 Online Magazines, (E): 531
 Real Data Reconstruction, (SA): 17
 Social Networking, (BE): 7, (E): 14, 422–423, 533, 534, 604, (IE): 2, 412, 445–446, (JC): 512, (SBS): 412–413
 Software, (E): 386, 559–560
 Speech Transcriptions, (E): 461–462

Stereograms, (E): 124–125, 608
 Telephone, (E): 6, 359
 Television, (E): 199, 278, 388, 424, 459,
 560, 604, 611–612, (IE): 258, (SBS):
 346–347
 Websites, (E): 456, 528, 557

Transportation

Accidents, (BE): 417–418, (E): 56, 59, 166,
 167–169, 461, 635, (IE): 414–415, 536
 Air Travel, (E): 59, 90, 115, 116–117, 204,
 280, 286, 390, 427, 428, 492, 532, 557,
 (IE): 403–404, 520, (JC): 104, 126
 Bicycle Safety, (E): 15, 171
 Bridge Safety, (E): 242, 285, 290, 387,
 (RM): 214–216, 433, 677, (SBS):
 213–214
 Bureau of Transportation Statistics,
 (JC): 104
 Car Manufacture, (E): 460, 668, 671
 Car Ownership, (E): 530

Car Purchases, (E): 459, 605
 Car Repairs, (E): 424, 529
 Car Speeds, (E): 156, 157, 282, (RM):
 591–592
 Commute Distances, (BE): 620–621,
 623–624
 Commute Times, (E): 387, (IE): 483, (JC):
 137, (RM): 43, 139–141, (SBS): 480–481
 Distance Traveled, (E): 668
 Drivers' Licenses, (BE): 505, 506, 507,
 (E): 95
 Drivers Type, (E): 427
 Driving Speeds, (E): 424, 568, (RM):
 105–106
 E-Bikes, (E): 15
 Emissions Testing, (E): 462, 559
 Engine Size, (E): 170
 Fuel Economy, (E): 16, 116, 120–121, 123,
 154, 201, 202, 234, 244–245, 286, 287,
 289, 327–328, 359, 383, 389, 633,
 637–638, 713, 714, (IE): 148–149,
 261–262, (JC): 39, (SBS): 40–41

Gas Prices, (E): 63, 120
 Gasoline Supply, (E): 331
 Horsepower, (E): 327
 Motorcycles, (IE): 536
 Roadblocks, (E): 357
 Sobriety Checkpoints, (E): 428
 Speeding, (E): 424, 568
 Stopping Distances, (E): 288, 496, 638
 Stopping Times, (E): 277
 Texting While Driving, (IE): 2
 Tires, (E): 159
 Traffic/Parking/Safety, (BE): 404, (E): 14,
 94, 200, 326, 491–492, 528, 557, 559,
 570, 634–635, (IE): 391–393, 571, 572,
 (JC): 662, (SBS): 575–576
 Trains, (E): 422, 425
 Used Cars, (E): 240, 241, 712, 713
 Weights of Vehicles, (E): 204, 235, 332, 633

This page intentionally left blank

SIXTH EDITION

Intro Stats

This page intentionally left blank

Stats Starts Here¹

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 the science of learning from 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.

1.1 What Is Statistics?

1.2 Data

1.3 Variables

1.4 Models

“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 heart 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 these 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 develop 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 Andrews McMeel Syndication. 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 *Wired* magazine,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, gender, 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.
- ◆ Americans spend an average of 4.9 hours per day on their smartphones. About 9.4 trillion text messages are sent each year.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 drivers who were above the legal limit for alcohol.

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

²<http://www.wired.com/story/wired-guide-personal-data-collection/>

³<https://www.textrequest.com/blog/texting-statistics-answer-questions/>

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

1.2 Data

STATISTICS IS ABOUT...

- ◆ Variation: Data vary because we don't see everything, and even what we do see, we measure imperfectly.
- ◆ Learning from data: We hope to learn about the world as best we can from the limited, imperfect data we have.
- ◆ Making intelligent decisions: The better we understand the world, the wiser our decisions will be.

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 2019, the company’s sales reached almost \$280.5 billion (more than 22% over 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”? You might think that data have to be numbers, but data can be text, pictures, web pages, and even audio and video. If you can sense it, you can measure it. The amount of data collected in the world is growing exponentially.⁵

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

B0000010AA	0.99	Chris G.	902	105-2686834-3759466	1.99	0.99	Illinois
Los Angeles	Samuel R.	Ohio	N	B000068ZVQ	Amsterdam	New York, New York	Katherine H.
Katherine H.	002-1663369-6638649	Beverly Hills	N	N	103-2628345-9238664	0.99	Massachusetts
312	Monique D.	105-9318443-4200264	413	B0000015Y6	440	B000002BK9	0.99
Canada	Detroit	440	105-1372500-0198646	N	B002MXA7Q0	Ohio	Y

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	Download	Gift?	ASIN	Artist
105-2686834-3759466	Katherine H.	Ohio	0.99	440	Amsterdam	N	B0000015Y6	Cold Play
105-9318443-4200264	Samuel R.	Illinois	1.99	312	Detroit	Y	B000002BK9	Red Hot Chili Peppers
105-1372500-0198646	Chris G.	Massachusetts	0.99	413	New York, New York	N	B000068ZVQ	Frank Sinatra
103-2628345-9238664	Monique D.	Canada	0.99	902	Los Angeles	N	B0000010AA	Blink 182
002-1663369-6638649	Katherine H.	Ohio	0.99	440	Beverly Hills	N	B002MXA7Q0	Weezer

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.

You should always stop to consider the ethical issues around collecting, managing, visualizing, and analyzing data. Throughout this text, we’ll present ethics discussions and examples. Because these are real examples, each is complex and has no one right solution. We hope they stimulate further discussion in and out of class.

⁵But not at a rate that researchers seem to be able to agree upon. It may be doubling every year or growing by as much as ten-fold every two years, depending on whom you believe.

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

DATA BEATS INTUITION

Amazon monitors and updates its website to better serve customers and maximize sales. To decide which changes to make, analysts experiment with new designs, offers, recommendations, and links. Statisticians want to know how long you'll spend browsing the site and whether you'll follow the links or purchase the suggested items. As Ronny Kohavi, former director of Data Mining and Personalization for Amazon, said, "Data trumps intuition. Instead of using our intuition, we experiment on the live site and let our customers tell us what works for them."

Look at all the columns to see exactly what each row refers to. Here the cases are different purchase records. You might have thought that each customer was a case, but notice that, for example, Katherine H. appears twice, in both 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.

ETHICS MATTERS

In the United States, the Belmont Report⁶ is the main federal document that provides the "Ethical Principles and Guidelines for the Protection of Human Subjects of Research."

The three fundamental ethical principles for using any human subjects for research are:

1. **Respect for persons:** The autonomy of all people should be protected. They should be treated with courtesy and respect and provided informed consent. Researchers must be truthful and conduct no deception.
2. **Beneficence:** The analyst must "do no harm" while maximizing benefits for the research project and minimizing risks to the research subjects.
3. **Justice:** There must be reasonable, nonexploitative, and well-considered procedures, administered fairly—a fair distribution of costs and benefits to potential research participants—and equally.

Respect for Persons

Data collection should respect a person's identity, including their gender identity. Although non-binary gender and gender fluidity are becoming more widely accepted in Western societies, diversity in gender identity is not new. Many indigenous cultures and other societies have recognized more than two genders throughout history. In the past few years gender fluidity has become increasingly important and prevalent. A 2016 Harris poll found that 1% of all millennials in the US identify as bigender. In that same year, Jamie Shupe became the first person in the US to be granted official non-binary gender status.

⁶<https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html>

However, government agencies worldwide have been slow to adapt their data collection procedures. An informal survey of agencies such as the United Nations, the World Health Organization, the 2020 US Census, the Centers for Disease Control shows that (at least until very recently) they collect gender data with only binary choices. We continue to use these data as collected, and they appear in this text in the way they are reported. Among the reasons for continuing to use them is that they are vital for the continuing study of the economic and health-related inequality experienced by women.

As new data are collected more gender options will appear. If you collect your own data, you should certainly take a more inclusive approach. But realize that there are many data sets with important information that currently have only binary choices and that this situation may only evolve slowly.

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 their 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.



EXAMPLE 1.1

Identifying the *Who*

Consumer Reports published an evaluation of 126 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 126 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.

“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”



Categorical Variables

Some variables just tell us what group or category each individual belongs to. Do you wear glasses or not? Are you 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.

Descriptive responses to questions are often categories. For example, the responses to the questions “Who is your cell phone provider?” and “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 next-day air. Counting is a natural way to summarize a categorical variable such as *Shipping Method*. Chapters 2 and 3 discuss 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 a quantitative variable, 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 Mauritanian Ouguiya (MUR).⁷

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 a.m. 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.

⁷As of 3/21/2020 \$1 = 37.32 MUR

Identifiers

For a categorical variable like *Survived*, each individual is assigned one of two possible values, say *Alive* or *Dead*⁸. 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. Identifiers are part of what's called **metadata**, or data about the data. Metadata 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.⁹ 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, and is needed so that the electronic messages sent to you can find you.

ETHICS MATTERS

You have many identifiers: a Social Security number, a student ID number, possibly a passport number, a health insurance number, and probably a Google account name. Privacy experts are worried that cyber thieves may match your identity in these different areas of your life, allowing, for example, your health, education, and financial records to be merged. 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. Did you realize that you are one of the cases in these data sets? Do you know what they are doing with your identifying data? The National Strategy for Trusted Identities in Cyberspace (www.wired.com/images_blogs/threatlevel/2011/04/NSTICstrategy_041511.pdf) proposes ways that we may address this challenge in the near future.

Ordinal Variables

A typical course evaluation survey asks, "How valuable do you think this course will be to you?" 1 = Worthless; 2 = Slightly; 3 = Middling; 4 = Reasonably; 5 = Invaluable. Is *Educational Value* categorical or quantitative? Often the best way to tell is to look to the *Why* of the study. A teacher might just count the number of students who gave each response for her course, treating *Educational Value* as a categorical variable. When she wants to see whether the course is improving, she might treat the responses as the *amount* of perceived value—in effect, treating the variable as quantitative.

But what are the units? There is certainly an *order* of perceived worth: Higher numbers indicate higher perceived worth. A course that averages 4.5 seems more valuable than one that averages 2, but we should be careful about treating *Educational Value* as purely quantitative. To treat it as quantitative, she'll have to imagine that it has "educational

⁸Well, maybe three values if you include Zombies.

⁹The National Security Agency (NSA) made the term "metadata" famous in 2014 by insisting that they only collected metadata on U.S. citizens' phone calls and text messages, not the calls and messages themselves. They later admitted to the bulk collection of actual data. In fact, some people say that the NSA is the only government agency that really listens to you.

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.

EXAMPLE 1.2

Identifying the *What* and *Why* of Tablets

RECAP: A *Consumer Reports* article about 126 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)
- quality score (quantitative, no units)
- memory card reader (categorical)

The magazine hopes to provide consumers with the information that will help them choose a good tablet.



JUST CHECKING

In the 2004 Tour de France bicycle race, 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 his titles; in addition, his records were expunged. You can find data on all the Tour de France races in the data set **Tour de France 2020**. Here are the first three and last nine 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
...										
2012	Bradley Wiggins	Great Britain	32	Sky	87.34.47	39.83	20	3488	198	153
2013	Christopher Froome	Great Britain	28	Sky	83.56.40	40.55	21	3404	198	169
2014	Vincenzo Nibali	Italy	29	Astana	89.56.06	40.74	21	3663.5	198	164
2015	Christopher Froome	Great Britain	30	Sky	84.46.14	39.64	21	3660.3	198	160
2016	Christopher Froome	Great Britain	31	Sky	89.04.48	39.62	21	3529	198	174
2017	Christopher Froome	Great Britain	32	Sky	86.34	40.997	21	3540	198	167
2018	Geraint Thomas	Great Britain	32	Sky	83.28	40.210	21	3349	176	145
2019	Egan Bernal	Colombia	22	INEOS	82.57.00	40.576	21	3365.8	176	155
2020	Tadej Pogacar	Slovenia	21	UAE Team Emirates	87.20.05	39.872	21	3482.2	176	146



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.

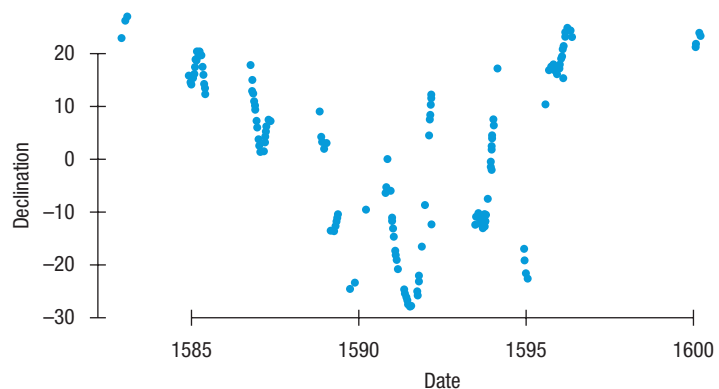
1.4 Models

What is a **model** for data? Models are summaries and simplifications of data that help our understanding in many ways. We'll encounter all sorts of models throughout the book. A model is a simplification of reality that gives us information that we can learn from and use, even though it doesn't represent reality exactly. A model of an airplane in a wind tunnel can give insights about the aerodynamics and flight performance of the plane even though it doesn't show every rivet.¹⁰ In fact, it's precisely because a model is a simplification that we learn from it. Without making models for how data vary, we'd be limited to reporting only what the data we have at hand say. To have an impact on science and society we'll have to generalize those findings to the world at large.

Kepler's laws describing the motion of planets are a great example of a model for data. Using astronomical observations of Tycho Brahe, Kepler saw through the small anomalies in the measurements and came up with three simple "laws"—or models for how the planets move. Here are Brahe's observations on the declination (angle of tilt to the sun) of Mars over a twenty-year period just before 1600:

Figure 1.1

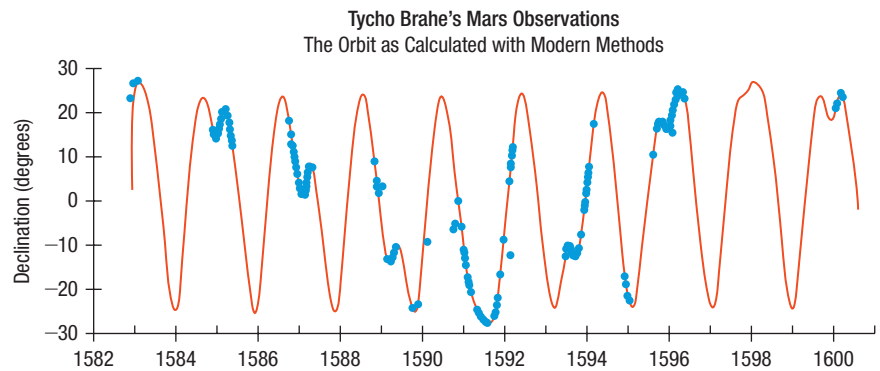
A plot of declination against time shows some patterns. There are many missing observations. Can you see the model that Kepler came up with from these data?



¹⁰Or tell you what movies you might see on the flight.

Figure 1.2

The model that Kepler proposed filled in many of the missing points and made the pattern much clearer.



Later, after Newton laid out the physics of gravity, it could be shown that the laws follow from other principles, but Kepler derived the models from data. We may not be able to come up with models as profound as Kepler's, but we'll use models throughout the book. We'll see examples of models as early as Chapter 5 and then put them to use more thoroughly later in the book when we discuss inference.

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.

CHAPTER REVIEW



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

- ◆ *Who, what, why, where, when* (and *how*)—the W's—help nail down the context of the data.
- ◆ We must know *who, what, and why* to be able to say anything useful based on the data. The *Who* are the cases. The *What* are the variables. A variable gives information about each of the cases. The *Why* helps us decide which way to treat the variables.
- ◆ Stop and identify the W's whenever you have data, and be sure you can identify the cases and the variables.

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

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

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

REVIEW OF TERMS

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

Data	Recorded values, whether numbers or labels, together with their context (p. 1).
Data table	An arrangement of data in which each row represents a case and each column represents a variable (p. 3).
Context	The context ideally tells <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. 3).
Case	An individual about whom or which we have data (p. 4).
Respondent	Someone who answers, or responds to, a survey (p. 4).
Subject	A human experimental unit. Also called a participant (p. 4).
Participant	A human experimental unit. Also called a subject (p. 4).
Experimental unit	An individual in a study for which or for whom data values are recorded. Human experimental units are usually called subjects or participants (p. 4).
Record	Information about an individual in a database (p. 4).
Sample	A subset of a population, examined in hope of learning about the population (p. 4).
Population	The entire group of individuals or instances about whom we hope to learn (p. 4).
Variable	A variable holds information about the same characteristic for many cases (p. 5).
Categorical (or qualitative) variable	A variable that names categories with words or numerals (p. 6).
Nominal variable	The term “nominal” can be applied to a variable whose values are used only to name categories (p. 6).
Quantitative variable	A variable in which the numbers are values of measured quantities with units (p. 6).
Units	A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams (p. 6).
Identifier variable	A categorical variable that records a unique value for each case, used to name or identify it (p. 7).
Ordinal variable	The term “ordinal” can be applied to a variable whose categorical values possess some kind of order (p. 8).
Model	A description or representation, in mathematical and statistical terms, of the behavior of a phenomenon based on data (p. 9).

TECH SUPPORT

Entering Data

These days, nobody does statistics by hand. We use technology: a programmable calculator or a statistics program on a computer. Professionals all use a *statistics package* designed for the purpose. We will provide many examples of results from a statistics package throughout the book. Rather than choosing one in particular, we'll offer generic results that look like those produced by all the major statistics packages but don't exactly match any of them. Then, in the Tech Support section at the end of each chapter, we'll provide hints for getting started on several of the major packages.

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 tab-delimited text (.txt) format.

DATA DESK

To read data into Data Desk:

- ▶ Click the **Open File** icon or choose **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.
- ▶ **File > Import** works 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.

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 can read 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 for data you plan to move to a statistics package.

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 website. **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.

Click a data set link to analyze the data or edit its properties to import a new data set.

- ▶ Choose **Select a file on my computer**,
- ▶ Enter the URL of a file,
- ▶ Paste data into a form, or
- ▶ Type or paste data into a blank data table.

For the “select a file on my computer” option, Statcrunch offers a choice of space, 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.

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?
3. **Parking lots** Sensors in parking lots are able to detect and communicate when spaces are filled in a large covered parking garage next to an urban shopping mall. How might the owners of the parking garage use this information both to attract customers and to help the store owners in the mall make business plans?
4. **Satellites and global climate change** Satellites send back nearly continuous data on the earth's land masses, oceans, and atmosphere from space. How might researchers use this information in both the short and long term to help study changes in the earth's climate?

SECTION 1.2

5. **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.
6. **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.
7. **Health records** The National Center for Health Statistics (NCHS) conducts an extensive survey consisting of an interview and medical examination with a representative sample of about 5000 people a year. The interview includes demographic, socioeconomic, dietary, and other health-related questions. The examination “consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel” (www.cdc.gov/nchs/nhanes/about_nhanes.htm). Describe the sample, the population, the *Who* and the *What* of this study.
8. **Facebook** Facebook uploads more than 350 million photos every day onto its servers. For this collection, describe the *Who* and the *What*.

SECTION 1.3

9. **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.
10. **ZIP codes** The U.S. 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 higher ZIP code tell you anything about a location compared to a lower ZIP code?
11. **Gay marriage** A May 2020 Gallup poll asked, “Do you think marriages between same-sex couples should or should not be recognized by the law as valid, with the same rights as traditional marriages?” “Sixty-seven percent of respondents said they should be valid—a new high in approval. If the choices were “Should be valid” and “Should not be valid”, what kind of variable is the response?
12. **Gay marriage by party** The May 2020 Gallup poll cited in Exercise 11 also differentiated respondents according to the political party they identified with. Gallup reports that 83% of Democrats responded that gay marriage should be recognized, but only 49% of those self-identifying as Republican did so. What kind of variable is the response percentages Gallup reports?
13. **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?
14. **Stress** A medical researcher measures the increase in heart rate of patients who are taking a stress test. What kind of variable is the researcher studying?

SECTION 1.4

15. **Voting and elections** Pollsters are interested in predicting the outcome of elections. Give an example of how they might model whether someone is likely to vote.
16. **Weather** Meteorologists utilize sophisticated models to predict the weather up to ten days in advance. Give an example of how they might assess their models.
17. **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.
18. **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 19–26) For each description of data, identify Who and What were investigated and the Population of interest.

19. **Gaydar** A study conducted by a team of American and Canadian researchers found that during ovulation, a woman can tell whether a man is gay or straight by looking at his face. To explore the subject, researchers recruited 40 undergraduate women who were not using any contraceptive drugs. These participants were tasked with guessing the sexual orientation of 80 men based on photos of their faces. The photographed men were all judged to be equally attractive and had neutral expressions. The study found that the closer a woman was to her peak ovulation, the more accurate were her guesses. <https://consumer.healthday.com/mental-health-information-25/behavior-health-news-56/does-ovulation-boost-a-woman-s-gaydar-654279.html>
20. **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/)
21. **Bicycle safety** Ian Walker, a psychologist at the University of Bath, wondered whether drivers treat bicycle riders differently when they wear helmets. He rigged his bicycle with an ultrasonic sensor that could measure how close each car was that passed him. He then rode on alternating days with and without a helmet. Out of 2500 cars passing him, he found that when he wore his helmet, motorists passed 3.35 inches closer to him, on average, than when his head was bare. (Source: *NY Times*)
22. **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.
23. **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*)
24. **Blindness** A study begun in 2011 examines the use of stem cells in treating two forms of blindness, Stargardt's disease and dry age-related macular degeneration. Each of the 24 patients entered one of two separate trials in which embryonic stem cells were to be used to treat the condition. (www.blindness.org/index.php?view=article&id=2514:stem-cell-clinical-trial-for-stargardt-disease-set-to-begin-&option=com_content&Itemid=122)
25. **Not-so-diet soda** A look at 474 participants in the San Antonio Longitudinal Study of Aging found that participants who drank two or more diet sodas a day "experienced waist size

increases six times greater than those of people who didn't drink diet soda." (*J Am Geriatr Soc.* 2015 Apr;63(4):708–15. doi: 10.1111/jgs.13376. Epub 2015 Mar 17.)

26. **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 27–40) 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).

27. **Weighing bears** Because of the difficulty of weighing a bear in the woods, researchers caught and measured 54 bears, recording their weight, neck size, length, and sex. They hoped to find a way to estimate weight from the other, more easily determined quantities.
28. **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.
29. **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.
30. **Religious Landscape.** The Pew Research Center 2014 Religious Landscape study surveyed a representative sample of U.S. Adults. Among the questions asked were the respondents' Age, Gender, Education, Marital Status, Belief in God (certain, fairly certain, not certain, don't know, do not believe in God), and Frequency of Prayer (daily, weekly, monthly, seldom/never).
31. **E-bikes.** In May of 2020 *Bicycling* magazine reviewed electric bicycles. For each bike reviewed, they reported the motor size (in watts), maximum speed (mph), wheel base (mm), brand name, and whether the battery can be removed for security. They also provided pictures of each bike.
32. **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.
33. **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.

Year	Winner	Jockey	Trainer	Owner	Time
1875	Aristides	O. Lewis	A. Williams	H. P. McGrath	2:37.75
1876	Vagrant	R. Swim	J. Williams	William Astor	2:38.25
1877	Baden Baden	W. Walker	E. Brown	Daniel Swigert	2:38
1878	Day Star	J. Carter	L. Paul	T. J. Nichols	2:37.25
...					
2013	Orb	J. Rosario	S. McGaughey	Stuart Janney & Phipps Stable	2:02.89
2014	California Chrome	Victor Espinoza	Art Sherman	California Chrome, LLC	2:03.66
2015	American Pharoah	Victor Espinoza	Bob Baffert	Zayat Stables, LLC	2:03.03
2016	Nyquist	M. Gutierrez	Doug F. O'Neill	Reddam Racing LLC	2:01.31
2017	Always Dreaming	J. Velazquez	Todd Pletcher	Meb Racing Stables	2:03.59
2018	Justify	M. Smith	Bob Baffert	China Horse Club	2:04.:20
2019	Country House	Flavien Prat	Bill Mott	J.O. Shields et al,	2:03.93
2020	Authentic	John Velazquez	Bob Baffert	Spendthrift Farm and others	2:00.61

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

34. Vineyards Business analysts hoping to provide information helpful to American grape growers compiled these data about vineyards: size (acres), number of years in existence, state, varieties of grapes grown, average case price, gross sales, and percent profit.

35. 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).

36. 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.

37. Dogs detecting coronavirus. In 2020, researchers at the University of Helsinki studied whether dogs can be trained to identify coronavirus (COVID-19) by smell. Researchers exposed dogs who had previously learned to identify cancer by scent to urine samples from individuals who had either tested positive or negative for COVID-19. They reported that the dogs were quick to learn the new scent. (Source: <https://www.helsinki.fi/en/news/health-news/the-finnish-covid-dogs-nose-knows>)

38. 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 gender, 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)

T 39. Kentucky Derby 2020 The Kentucky Derby is a horse race that has been run every year since 1875 at Churchill Downs in 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 two—1901—when it took place on April 29, and 2020, when it was held on September 5.) The table at the top of the page shows the data for the first four and eight recent races.

T 40. Indy 500 2020 The 2.5-mile Indianapolis Motor Speedway has been the home to a race on Memorial Day nearly every year since 1911. Even during the first race, there were controversies. Ralph Mulford was given the checkered flag first but took three extra laps just to make sure he'd completed 500 miles. When he finished, another driver, Ray Harroun, was being presented with the winner's trophy, and Mulford's protests were ignored. Harroun averaged 74.6 mph for the 500 miles. In 2013, the winner, Tony Kanaan, averaged over 187 mph, beating the previous record by over 17 mph!

Here are the data for the first five races and five recent Indianapolis 500 races.

Year	Driver	Time (hr:min:sec)	Speed (mph)
1911	Ray Harroun	6:42:08	74.602
1912	Joe Dawson	6:21:06	78.719
1913	Jules Goux	6:35:05	75.933
1914	René Thomas	6:03:45	82.474
1915	Ralph DePalma	5:33:55.51	89.840
...			
2015	Juan Pablo Montoya	3:05:56.5286	161.341
2016	Alexander Rossi	3:00:02.0872	166.634
2017	Takuma Sato	3:13:3.3584	115.395
2018	Will Power	2:59:42.6365	166.935
2019	Simon Pagenaud	2:50:39.27948	175.79362
2020	Takuma Sato	3:10:05.0880	157.824

- T 41. Kentucky Derby 2020 on the computer** Load the Kentucky Derby data into your preferred statistics package and answer the following questions;
- What was the name of the winning horse in 1880?
 - When did the length of the race change?
 - What was the winning time in 1974?
 - Only one horse has run the Derby in less than 2 minutes. Which horse and in what year?
- T 42. Indy 500 2020 on the computer** Load the Indy 500 2020 data into your preferred statistics package and answer the following questions:
- What was the average speed of the winner in 1920?
 - How many times did Bill Vukovich win the race in the 1950s?
 - How many races took place during the 1940s?

JUST CHECKING

Answers

1. *Who*—Tour de France races; *What*—year, winner, country of origin, age, team, total time, average speed, stages, total distance ridden, starting riders, finishing riders; *How*—official statistics at race; *Where*—France (for the most part); *When*—1903 to 2019; *Why*—not specified (To see progress in speeds of cycling racing?)

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

STUDENT ACTIVITY

Real Data Reconstruction

Throughout this course, you will often start with a dataset and then work to analyze it. Let's reverse that process and start with a graphic and think about the original data that was needed to produce it.

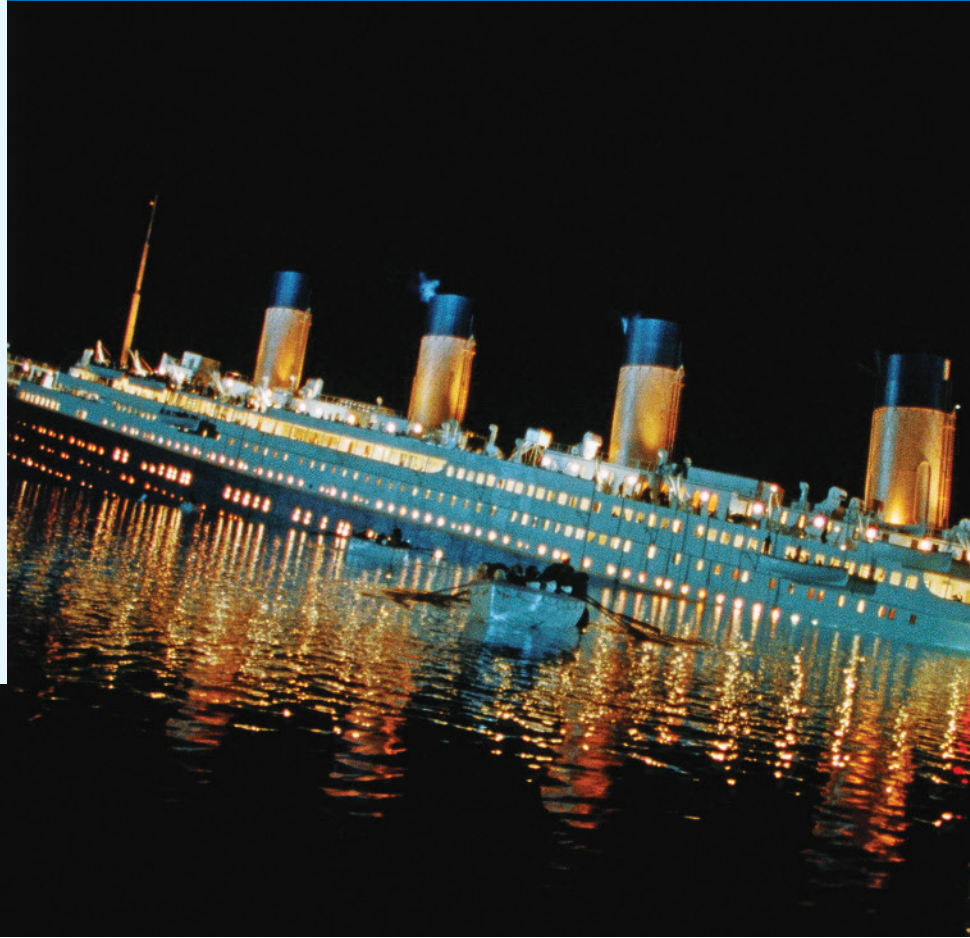
- ◆ **Step 1.** Find a visualization of data from a reliable news source, available online or in print.
- ◆ **Step 2.** Based on the article or surrounding information, detail the full context of the data used to create the visualization. In particular, list **how** the data were collected, **where** they were collected, and **when** and **why** they were collected. Make it clear which of these are provided by the news source. If some context is unavailable, make a guess but indicate that as such.
- ◆ **Step 3.** Now, let's reconstruct the original data set; consider who was measured (rows) and what was measured (columns). Visualizations provide summaries, so you must consider who or what one **case** represents and what **variable(s)** or characteristic(s) are recorded. Based on the information provided by the news source, detail what you think the data set looks like by describing the cases and variables.
- ◆ **Step 4.** Lastly, consider the variables or characteristics represented in the visualization. Write a brief paragraph describing whether the variables are being treated as **quantitative** variables or **categorical** variables. Remember that the visualization may show a percentage of individuals with a particular characteristic so that you have a numerical summary of a categorical variable.

2

Displaying and Describing Data

WHERE ARE WE GOING?

We can summarize and describe data values in a variety of ways. You'll probably recognize these displays and summaries. This chapter is a fast review of these concepts so we all agree on terms, notation, and methods. We'll be using these displays and descriptions throughout the rest of the book.



- 2.1 Summarizing and Displaying a Categorical Variable
- 2.2 Displaying a Quantitative Variable
- 2.3 Shape
- 2.4 Center
- 2.5 Spread

What happened on the *Titanic* at 11:40 on the night of April 14, 1912, is well known. Frederick Fleet's cry of "Iceberg, right ahead" and the three accompanying pulls of the crow's nest bell signaled the beginning of a nightmare that has become legend. By 2:15 a.m., the *Titanic*, thought by many to be unsinkable, had sunk. Only 712 of the 2208 people on board survived. The others (nearly 1500) met their icy fate in the cold waters of the North Atlantic.

Table 2.1 shows some data about the passengers and crew aboard the *Titanic*. Each case (row) of the data table represents a person on board the ship. The variables are the person's *Name*, *Survival* status (Lost or Saved), *Age* (in years), *Age Category* (Adult or Child), *Sex* (Male or Female), *Price Paid* (in British pounds, £), and ticket *Class* (First, Second, Third, or Crew). Some of these, such as *Age* and *Price*, record numbers. These are

Table 2.1

Part of a data table showing seven variables for 11 people aboard the *Titanic*.

Name	Survived	Age	Gender	Price	Class
ABBING, Mr Anthony	LOST	41	Male	7.55	3rd
ABBOTT, Mr Eugene Joseph	LOST	13	Male	20.25	3rd
ABBOTT, Mr Rossmore Edward	LOST	16	Male	20.25	3rd
ABBOTT, Mrs Rhoda Mary "Rosa"	SAVED	39	Female	20.25	3rd
ABELSETH, Miss Kalle (Karen) Marie Kristiane	SAVED	16	Female	7.65	3rd
ABELSETH, Mr Olaus Jørgensen	SAVED	25	Male	7.65	3rd
ABELSON, Mr Samuel	LOST	30	Male	24	2nd
ABELSON, Mrs Anna	SAVED	28	Female	24	2nd
ABİ SA'B, Mr Jirjis Yūsuf	LOST	45	Male	7.23	3rd
ABİ SA'B, Mrs Sha'ninah	SAVED	38	Female	7.23	3rd
ABİ SHADİD, Mr Dāhir	LOST	19	Male	7.23	3rd

WHO	People on the <i>Titanic</i>
WHAT	Name, survival status, age, adult/child, sex, price paid, ticket class
WHEN	April 14, 1912
WHERE	North Atlantic
HOW	www.encyclopedia-titanica.org
WHY	Historical interest

Titanic 2020?

Why would we put a date of 2020 on data about an event that occurred in 1912? Well, although no one who sailed on the *Titanic* is still alive, the dataset is very much alive. The site *Encyclopedia Titanica* has been updating the data and collecting background details about the lives of the people on board since it first went online in 1996. Relatives of those on board and history buffs have been adding and correcting information continuously. One of our authors has found mistakes in the data using basic statistical methods you'll learn in this text. Those errors have been corrected at the site and in our data. Sometimes a simple display can reveal a data value that can't be right. The data in **Titanic 2020** are the best we have as of the publication of this book. Perhaps you'll make more corrections!

called **quantitative** variables. Others, like *Survival* and *Class*, place each case in a single category, and are called **categorical** variables. (Data in **Titanic 2020**)

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

The Three Rules of Data Analysis

There are three things you should always do first with data:

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

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

We make graphs for two primary reasons: to understand more about data and to show others what we have learned and want them to understand. The first reason calls for simple graphs with little adornment; the second often uses visually appealing additions to draw the viewer's attention. Regardless of their function, graphs should be easy to read and understand and should represent the facts of the data honestly. Axes should be clearly labeled with the names of the variables they display. The intervals set off by “tick marks” should occur at values that are easy to think about: 5, 10, 15, and 20 are simpler marks than, say, 1.7, 2.3, 2.9, and 3.5. And tick labels that run for several digits are almost never a good idea. Graphs should have a “key” that identifies colors and symbols if those are meaningful in the graph. And all graphs should carry a title or caption that says what the graph displays and suggests what about it is salient or important.

The Area Principle

A bad picture can distort our understanding rather than help it. What impression do you get from Figure 2.1 about who was aboard the ship?

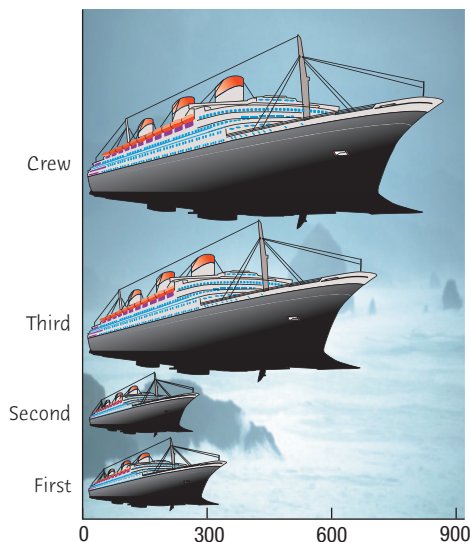


Figure 2.1

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

The *Titanic* was certainly a luxurious ship, especially for those in first class, but Figure 2.1 gives the mistaken impression that most of the people on the *Titanic* were crew members, with a few passengers along for the ride. What’s wrong? The lengths of the ships *do* match the number of people in each ticket class category. However, our eyes tend to be more impressed by the *area* than by other aspects of each ship image. So, even though the *length* of each ship matches up with one of the totals, it’s the associated *area* in the image that we notice. There were about 3 times as many crew as second-class passengers, and the ship depicting the number of crew members is about 3 times longer than the ship depicting second-class passengers. The problem is that it occupies about 9 times the area. That just isn’t a correct impression.

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

2.1 Summarizing and Displaying a Categorical Variable

Frequency Tables

Categorical variables are easy to summarize in a **frequency table** that lists the number of cases in each category along with its name.

For ticket *Class*, the categories are First, Second, Third, and Crew:

Class	Count
First	324
Second	284
Third	709
Crew	891

Table 2.2
A frequency table of the *Titanic* passengers.

Class	Percentage (%)
First	14.67
Second	12.86
Third	32.11
Crew	40.35

Table 2.3
A relative frequency table for the same data.

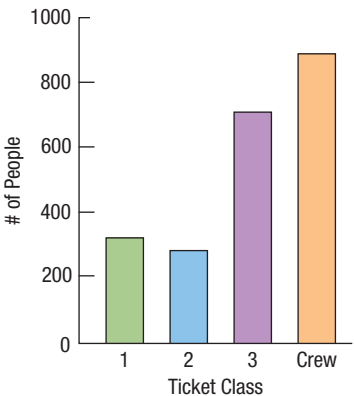


Figure 2.2
People on the Titanic by Ticket Class. With the area principle satisfied, we can see the true distribution more clearly.

A **relative frequency table** displays *percentages* (or *proportions*) rather than the counts in each category. Both types of tables show the **distribution** of a categorical variable because they name the possible categories and tell how frequently each occurs. (The percentages should total 100%, although the sum may be a bit too high or low if the individual category percentages have been rounded.)

Bar Charts

Although not as visually entertaining as the ships in Figure 2.1, the **bar chart** in Figure 2.2 gives an *accurate* visual impression of the distribution because it obeys the area principle. Now it’s easy to see that the majority of people on board were *not* crew. We can also see that there were about 3 times as many crew members as second-class passengers. And there were more than twice as many third-class passengers as either first- or second-class passengers—something you may have missed in the frequency table. Bar charts make these kinds of comparisons easy and natural.

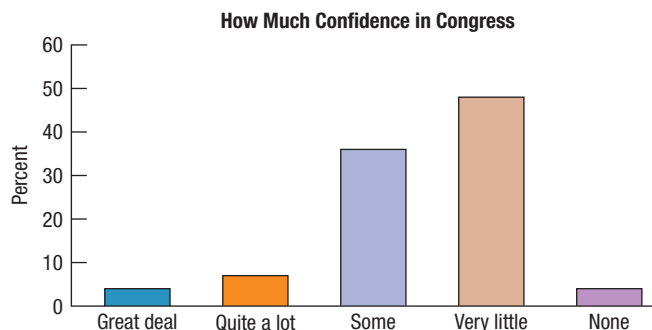
EXAMPLE 2.1**Confidence in Congress**

In June 2019 the Gallup poll asked a representative sample of U.S. adults how much confidence they had in various institutions in American society. The response about Congress was: (<https://news.gallup.com/poll/1600/congress-public.aspx>)

Great deal	4%
Quite a lot	7%
Some	36%
Very little	48%
None	4%

QUESTION: What kind of table is this? What would be an appropriate display?

ANSWER: This is a relative frequency table because the numbers displayed are percentages, not counts. A bar chart would be appropriate:

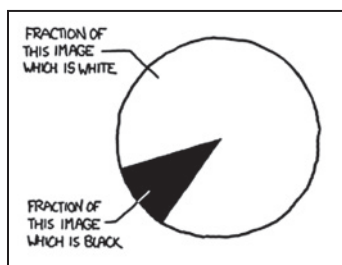
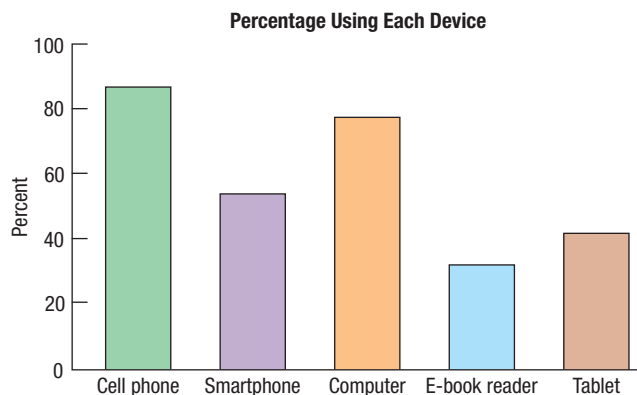
**EXAMPLE 2.2****Which Gadgets Do You Use?**

In 2014, the Pew Research Organization asked 1005 U.S. adults which of the following electronic items they use: cell phone, smartphone, computer, handheld e-book reader (e.g., Kindle or Nook), or tablet. The results were

Device	Percentage (%) using the device
Cell phone	86.8
Smartphone	54.0
Computer	77.5
E-book reader	32.2
Tablet	41.9

QUESTION: Is this a frequency table, a relative frequency table, or neither? How could you display these data?

ANSWER: This is not a frequency table because the numbers displayed are not counts. Although the numbers are percentages, they do not sum to 100%. A person can use more than one device, so this is not a relative frequency table either. A bar chart might still be appropriate, but the numbers do not sum to 100%.



© 2013 Randall Munroe. Reprinted with permission. All rights reserved.

Pie Charts

Pie charts display all the cases as a circle whose slices have areas proportional to each category's fraction of the whole.

Pie charts give a quick impression of the distribution. Because we're used to cutting up pies into 2, 4, or 8 pieces, pie charts are particularly good for seeing relative frequencies near $1/2$, $1/4$, or $1/8$.

Bar charts are almost always better than pie charts for comparing the relative frequencies of categories. Pie charts are widely understood and colorful, and they often appear in reports, but Figure 2.3 shows why statisticians prefer bar charts.

Figure 2.3

Pie charts may be attractive, but it can be hard to see patterns in them. Can you discern the differences in distributions depicted by these pie charts?

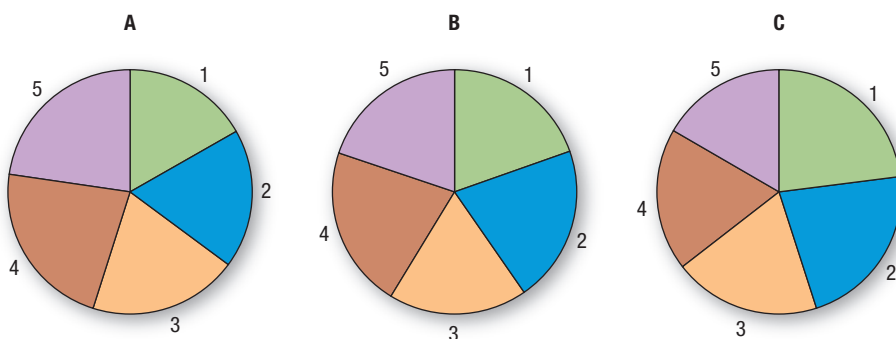
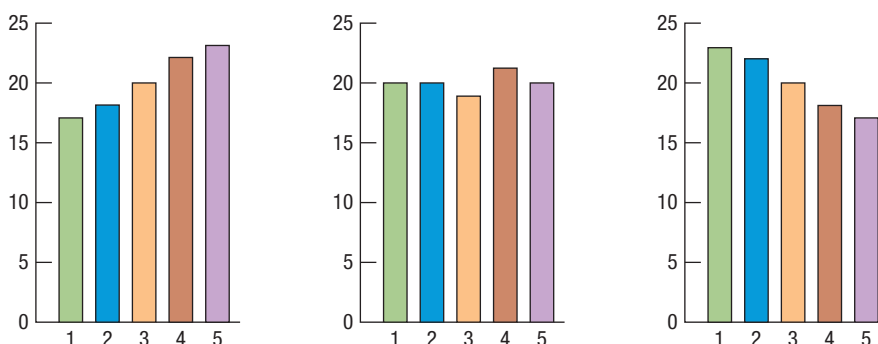


Figure 2.4

Bar charts of the same values as shown in Figure 2.3 make it much easier to compare frequencies in groups.



Ring Charts

A ring (or donut) chart is a modified form of pie chart that displays only the “crust” of the pie—a ring that is partitioned into regions proportional in area to each value. You can think of the ring as the bars of a bar chart stuck end to end and wrapped around the circle. Ring charts are somewhere between bar charts and pie charts. They may be easier to read (or not). Judge for yourself:

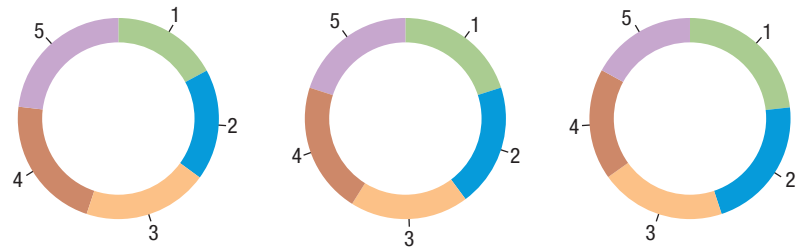


Figure 2.5

Ring charts compromise between pie and bar charts. These ring charts show the same values as the pie charts in Figure 2.3. Do you find it easier to see the patterns?



RANDOM MATTERS

Is it random, or is something systematic going on? Separating the *signal* (the systematic) from the *noise* (the random) is a fundamental skill of statistics.

A geoscientist notices that global temperatures have increased steadily during the past 50 years. Could the pattern be random, or is the earth warming?

An analyst notices that the stock market seems to go up more often on Tuesday afternoons when it rains in Chicago. Is that something she should bank on?

One of the challenges to answering questions like these is that we have only one earth and one stock market history. What if we had two? Or many? Sometimes we can use a computer to *simulate* other situations, to pretend that we have more than one realization of a phenomenon. In these *Random Matters* sections, we'll use the computer as our lab to test what might happen if we could repeat our data collection many times.

You probably know that the “rules of the sea” were enforced on the *Titanic*—women and children were allowed to board the *Titanic* lifeboats before the men. Did ticket class (first, second, or third) also make a difference? Suppose the 712 survivors were chosen at random, giving everyone an equal chance to get into a lifeboat. Would the distribution have been different? Let's look. We selected 712 people at random from the list of those aboard the *Titanic* and made a pie chart of ticket class. We repeated the random selection 24 times, making a new pie chart of each selected group of 712 passengers. Among these pie charts in Figure 2.6 we've “hidden” the actual distribution of survivors. Can you pick out the real distribution? If so, then that might convince you that the lifeboats weren't filled randomly, with everyone getting an equal chance.