SIXTH EDITION

Intro Stats

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.

Pearson

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, Rimpy 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

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 on page A-47, which constitutes an extension of this copyright page.

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://lccn.loc.gov/2021002385

ScoutAutomatedPrintCode

Pearson

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 capella* 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.

TABLE OF CONTENTS

*Indicates optional sections.

PREFACE

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

[1]And 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. Examples that use these datasets cite them in the text. More than 700 of our exercises have a T 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 T 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**

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^{TM}

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

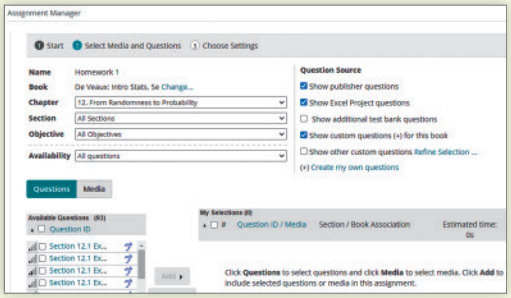
Resources for Success

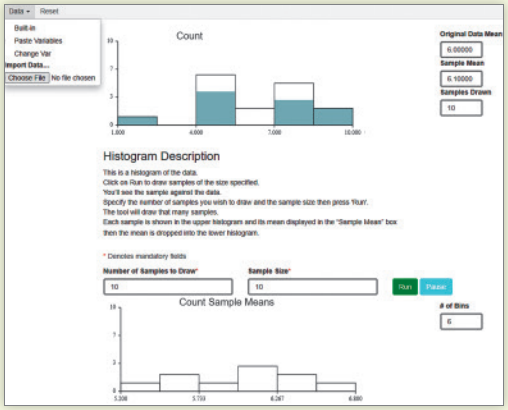
MyLab® Statistics Online Course for *Intro Stats, 6e* by Richard D. De Veaux, Paul F. Velleman, and David E. Bock (access code required)

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.

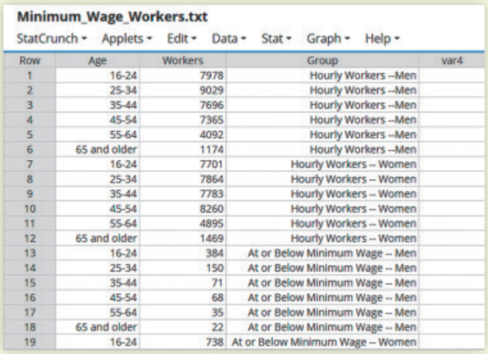
MyLab Statistics exercises have been updated to include real data so students can understand the real-world implications of data analysis. Homework reinforces and supports students' understanding of key statistics topics within a real world context.





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	

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.

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.

pearson.com/mylab/statistics

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
Cornell College

Susan Chimiak
University of Maryland

Lynda Hollingsworth
Northwest Missouri State University

Jeff Kollath
Oregon State University

Cindy Leary
University of Montana

Sheldon Lee
Viterbo University

Pam Omer
Western New England University

Sarah Quesen
West Virginia University

Karin Reinhold
SUNY Albany

Laura Shick
Clemson University

Dirk Tempelaar
Maastricht University

Carol Weideman
St. Petersburg College

Ming Wang
University of Kansas

Lisa Wellinghoff
Wright State

Cathy Zucco-Teveloff
Rider University

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

Mary Kay Abbey
Montgomery College

Froozan Pourboghnaf Afiat
Community College of Southern Nevada

Mehdi Afiat
Community College of Southern Nevada

Nazanin Azarnia
Santa Fe Community College

Sanjib Basu
Northern Illinois University

Carl D. Bodenschatz
University of Pittsburgh

Steven Bogart
Shoreline Community College

Ann Cannon
Cornell College

Robert L. Carson
Hagerstown Community College

Jerry Chen
Suffolk County Community College

Rick Denman
Southwestern University

Jeffrey Eldridge
Edmonds Community College

Karen Estes
St. Petersburg Junior College

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

Ken Grace
Anoka-Ramsey Community College

Jonathan Graham
University of Montana

Nancy Heckman
University of British Columbia

James Helreich
Marist College

Susan Herring
Sonoma State University

Mary R. Hudachek-Buswell
Clayton State University

Patricia Humphrey
Georgia Southern University

Becky Hurley
Rockingham Community College

Debra Ingram
Arkansas State University

Joseph Kupresanin
Cecil College

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

Energy

Environment

Famous People

Finance and Investments

Food/Drink

Games

Government, Labor, and Law

Human Resource Management/Personnel

Insurance

Manufacturing

Marketing

Media and Entertainment

Pharmaceuticals, Medicine, and Health

Politics and Popular Culture

Quality Control

Real Estate

Salary and Benefits

Sales and Retail

Science

Service Industries and Social Issues

Sports

Surveys and Opinion Polls

Technology

Transportation

SIXTH EDITION

Intro Stats

This page intentionally left blank

1

Stats Starts Here[1]

WHERE ARE WE GOING?

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

But statistics *is* fun. That's probably not what you heard on the street, but it's true. Statistics is 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.

"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

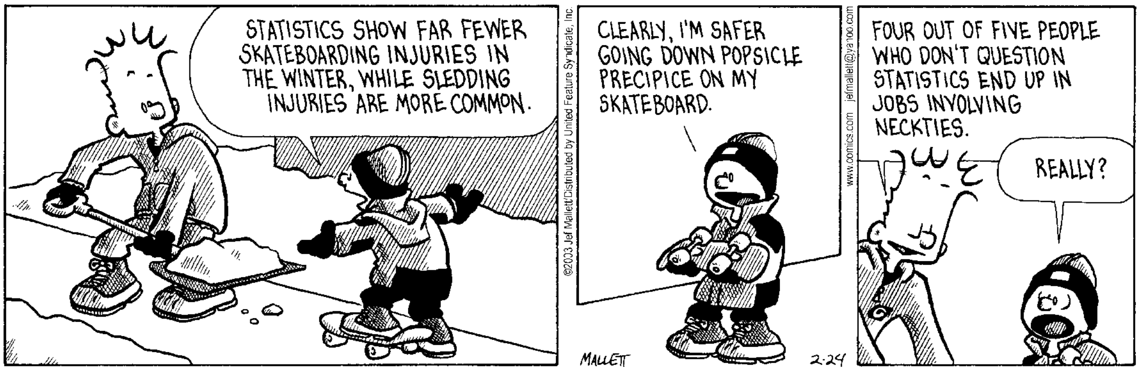
1.1 What Is Statistics?

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.

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



> "Data is king at Amazon. Clickstream and purchase data are the crown jewels at Amazon. They help us build features to personalize the Web site experience."
>
> —Ronny Kohavi, former Director of Data Mining and Personalization, Amazon.com

Q: What is statistics?

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

Q: What are statistics?

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

Q: So what is data?

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

Q: OK, OK, so what are data?

A: Data are values along with their context.

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

We say, "Don't be a datum."

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

Consider the following:

- If you have a Facebook account, you have probably noticed that the ads you see online tend to match your interests and activities. Coincidence? Hardly. According to *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	B00000I5Y6	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.

[5] But 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 Report[6] 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 identify, 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.

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.)

"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"

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.

[7] As 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 indentifying 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

[8]Well, maybe three values if you include Zombies.

[9]The 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 goverment 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:

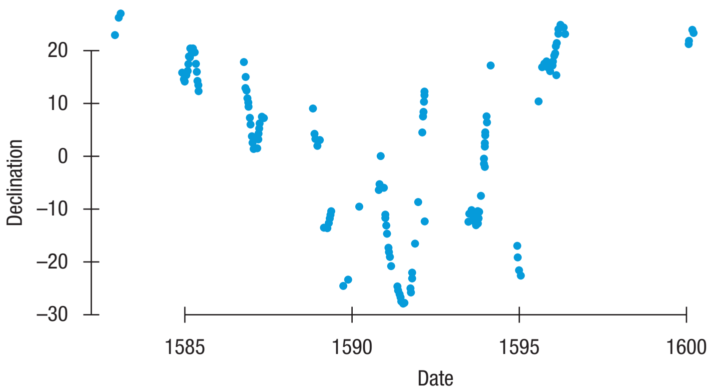


Figure 1.1
A plot of declination against time shows some patterns. There are many missing observations. Can you see the model that Kepler came up with from these data?

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