GLOBAL EDITION

Stats

Data and Models

FIFTH EDITION

De Veaux • Velleman • Bock

Get the Most Out of MyLab Statistics

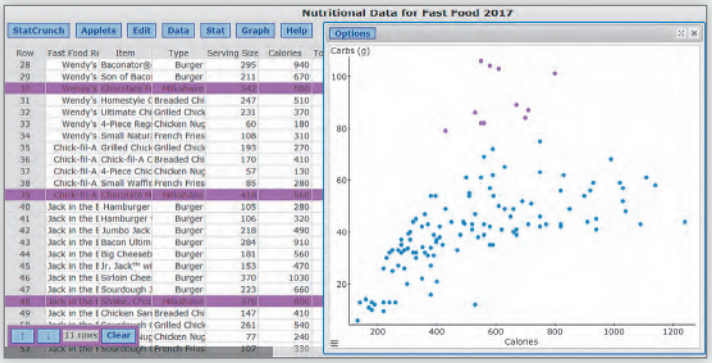
Pearson MyLab

MyLab™ Statistics is the teaching and learning platform that empowers instructors to reach *every* student. By combining trusted author content with digital tools and a flexible platform, MyLab Statistics personalizes the learning experience and improves results for each student.

Collect, crunch, and communicate with StatCrunch

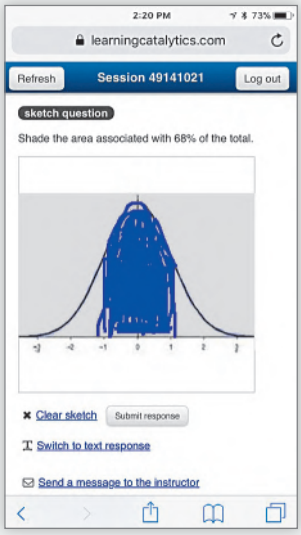
With StatCrunch®, Pearson's powerful web-based statistical software, instructors and students can access tens of thousands of data sets including those from the textbook, perform complex analyses, and generate compelling reports. StatCrunch is integrated directly into MyLab Statistics.

Beyond StatCrunch, MyLab Statistics makes learning and using a variety of statistical software packages seamless and intuitive by allowing users to download and copy data sets directly into other programs. Students can access instructional tools including tutorial videos, and study cards.



Give every student a voice with Learning Catalytics

Learning Catalytics™ is an interactive classroom tool that allows every student to participate. Instructors ask a variety of questions that help students recall ideas, apply concepts, and develop critical-thinking skills. Students answer using their smartphones, tablets, or laptops to show that they do—or don't—understand. Instructors monitor responses to adjust their teaching approach, and even set up peer-to-peer learning. More importantly, they use real-time analytics to address student misconceptions the moment they occur and ensure they hear from every student when it matters most.



Visit pearsonmylabandmastering.com and click Training & Support to make sure you're getting the most out of MyLab Statistics.

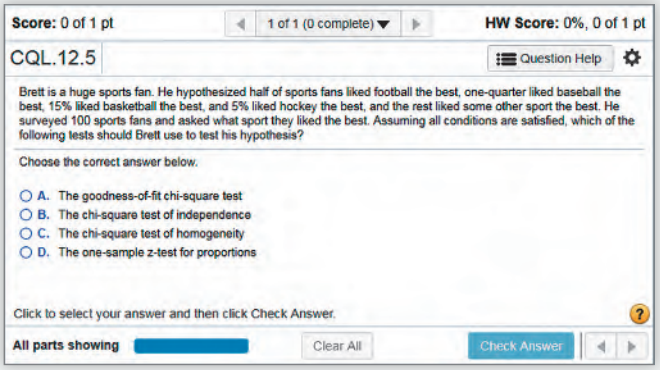
Pearson MyLab

Enrich assignments with question libraries

MyLab Statistics includes a number of question libraries providing additional opportunities for students to practice statistical thinking.

- StatCrunch Projects provide students with opportunities to analyze and interpret data. Each project consists of a series of questions about a large data set in StatCrunch.

- The Conceptual Question Library offers 1,000 conceptual-based questions to help students internalize concepts, make interpretations, and think critically about statistics.



- The Getting Ready for Statistics Library contains more than 450 exercises on prerequisite topics. Assign these questions to students who may need a little extra practice on their prerequisite skills to be successful in your course.

- The StatTalk Video Library is based on a series of 24 videos, hosted by fun-loving statistician Andrew Vickers, that demonstrate important statistical concepts through interesting stories and real-life events.

Incorporate additional author-created resources in class

Authors infuse their own voice, approach, and experiences teaching statistics into additional text-specific resources, such as interactive applets, technology manuals, workbooks, and more. Check out the Preface to learn more about what's available for this specific title.

This page is intentionally left blank

FIFTH EDITION

Stats: Data and Models

GLOBAL EDITION

Richard D. De Veaux
Williams College

Paul F. Velleman
Cornell University (Emeritus)

David E. Bock
Cornell University
Ithaca High School (Retired)

Pearson

Pearson Education Limited

KAO Two
KAO Park
Hockham Way
Harlow
CM17 9SR
United Kingdom

and Associated Companies throughout the world

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

Authorized adaptation from the United States edition, entitled Stats: Data and Models*, 5th Edition, ISBN 978-0-13-516382-5 by Richard D. De Veaux, Paul F. Velleman, and David E. Bock, published by Pearson Education © 2020.*

ISBN 10: 1-292-36221-9
ISBN 13: 978-1-292-36221-2
eBook ISBN 13: 978-1-292-36232-8

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Typeset by SPi Global
eBook formatted by B2R Technologies Pvt. Ltd.

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 (Ph.D. Stanford University) is an internationally known educator, consultant, and lecturer. Dick has taught statistics at a business school (Wharton), an engineering school (Princeton), and a liberal arts college (Williams). While at Princeton, he won a Lifetime Award for Dedication and Excellence in Teaching. Since 1994, he has taught at Williams College, although he returned to Princeton for the academic year 2006–2007 as the William R. Kenan Jr. Visiting Professor of Distinguished Teaching. He is currently the C. Carlisle and Margaret Tippit Professor of Statistics at Williams College. Dick holds degrees from Princeton University in Civil Engineering and Mathematics and from Stanford University, where he studied statistics with Persi Diaconis and dance with Inga Weiss. His research focuses on the analysis of large datasets and data mining in science and industry. Dick has won both the Wilcoxon and Shewell awards from the American Society for Quality. He is an elected member of the International Statistics Institute (ISI) and a Fellow of the American Statistical Association (ASA). Dick was elected Vice President of the ASA in 2018 and will serve from 2019 to 2021. Dick is also well known in industry, having consulted for such *Fortune* 500 companies as American Express, Hewlett-Packard, Alcoa, DuPont, Pillsbury, General Electric, and Chemical Bank. He was named the "Statistician of the Year" for 2008 by the Boston Chapter of the American Statistical Association. In his spare time he is an avid cyclist and swimmer, and is a frequent singer and soloist with various local choirs, including the Choeur Vittoria of Paris, France. Dick is the father of four children.

Paul F. Velleman (Ph.D. Princeton University) has an international reputation for innovative statistics education. He designed the Data Desk® software package and is also the author and designer of the award-winning ActivStats® multimedia software, for which he received the EDUCOM Medal for innovative uses of computers in teaching statistics and the ICTCM Award for Innovation in Using Technology in College Mathematics. He is the founder and CEO of Data Description, Inc. (**www.datadesk.com**), which supports both of these programs. Data Description also developed and maintains the Internet site *Data and Story Library* (DASL; **dasl.datadescription.com**), which provides many of the datasets used in this text as well as many others useful for teaching statistics, and the statistics conceptual tools at **astools.datadesk.com**. Paul coauthored (with David Hoaglin) the book *ABCs of Exploratory Data Analysis*. Paul is Emeritus Professor of Statistical Sciences at Cornell University, where he was awarded the MacIntyre Prize for Exemplary Teaching. Paul earned his M.S. and Ph.D. from Princeton University, where he studied with John Tukey. His research often focuses on statistical graphics and data analysis methods. Paul is a Fellow of the American Statistical Association and of the American Association for the Advancement of Science. He was a member of the working group that developed the GAISE 2016 guidelines for teaching statistics.

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 and a Statistics consultant to the College Board, leading workshops and institutes for AP Statistics teachers. His understanding of how students learn informs much of this book's approach.

Richard De Veaux, Paul Velleman, and David Bock have authored several successful books in the introductory college and AP High School market including *Intro Stats*, Fifth Edition (Pearson, 2018) and *Stats: Modeling the World*, Fifth Edition (Pearson, 2019).

*Indicates optional sections.

PART V Inference for One Parameter

PART VI Inference for Relationships

Stats: Data and Models, fifth edition, has been especially exciting to develop. The book you hold steps beyond our previous editions in several important ways. Of course, we've kept our conversational style and anecdotes,[1] but we've enriched that material with tools for teaching about randomness, sampling distribution models, and inference throughout the book. And we've expanded discussions of models for data to introduce models with more than two variables earlier in the text. 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, to expand our attention beyond just one or two variables, we've adjusted the order of some topics.

Innovations

Technology

One of the new GAISE guidelines states: *Use technology to explore concepts and analyze data.* 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 datasets. 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 *Stats: Data and Models*, we've taken this principle still further. We have harnessed technology 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. This edition of *Stats: Data and Models* introduces a third variable as early as Chapter 3's discussion of contingency tables and mosaic plots. Then, 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 early in the course 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 Chapters 23 and 24 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 until the chapter on ANOVA.)

[1] And footnotes

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 follows our presentation naturally. In this edition, we introduce a new feature, **Random Matters**. Random Matters elements in early chapters 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 a sample? The bootstrap provides a way to continue this line of thought, now by resampling 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. *Stats: Data and Models* 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. *Stats: Data and Models* comes with an archive of more than 500 datasets used in more than 700 applications throughout the book. The datasets are available online at pearsonglobaleditions.com and in MyLab Statistics. Examples that use these datasets cite them in the text. Exercises are marked when they use one of them; exercise names usually indicate the name of the dataset. We encourage students to get the datasets and reproduce our examples using their statistics software, and some of the exercises require that.

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 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 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 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.

Part IV provides background material on probability, random variables, and probability models. In Part V, Chapter 16 introduces confidence intervals for proportions as soon as we've reassured students that their intuition about the sampling distribution of proportions is correct. Chapter 17 formalizes the Central Limit Theorem and introduces Student's *t*-models. Chapter 18 is then about testing hypotheses, and Chapter 19 elaborates further, discussing alpha levels, Type I and Type II errors, power, and effect size. The chapters in Part VI deal with comparing groups (both with proportions and with means), paired samples, chi-square. Finally, Part VII discusses inferences for regression models (both simple and multiple), intelligent uses of multiple regression, and Analysis of Variance, both one- and two-way. A final chapter on data mining looks to the future.

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. The final set are "book-level" review problems that 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 a 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 24.)
- ***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 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 datasets, 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 book's website at pearsonglobaleditions.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.

New! Random Matters. This new feature travels 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).

Margin and in-text boxed notes. Throughout each chapter, boxed margin and in-text notes enhance and enrich the text.

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.

- **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.
- **Connections** specifically ties the new topics to those learned in previous chapters.
- 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 expanded and 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. First, you'll find relatively simple, focused exercises 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.

More than 600 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 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 seven 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-VII Cumulative Review Exercises. 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 book's website at **pearsonglobaleditions.com**
- Within the MyLab Statistics course at **pearsonmylabandmastering.com**

New tools that provide interactive versions of the distribution tables at the back of the book and tools for randomization inference methods such as the bootstrap and for repeated sampling from larger populations can be found online at **astools.datadesk.com**.

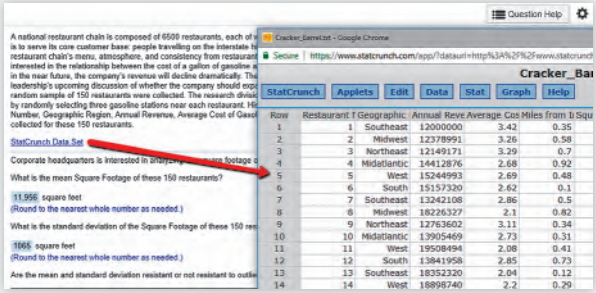
MyLab Statistics for *Stats: Data & Models, 5e*

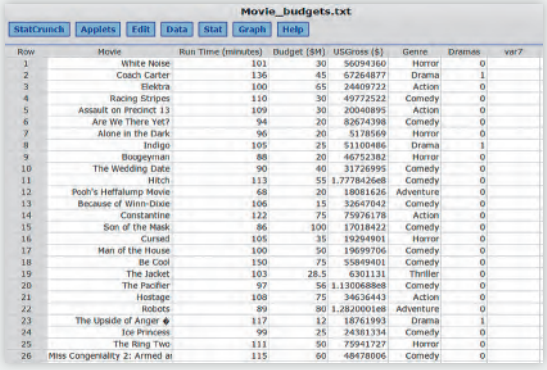
(access code required)

MyLab Statistics is available to accompany Pearson's market-leading text offerings. To give students a consistent tone, voice, and teaching method, each text's flavor and approach are tightly integrated throughout the accompanying MyLab course, making learning the material as seamless as possible.

NEW! StatCrunch Projects

StatCrunch Projects provide opportunities for students to explore data beyond the classroom. In each project, students analyze a data set in StatCrunch® and answer assignable MyLab questions for immediate feedback. StatCrunch Projects span the entire curriculum or focus on certain key concepts. Questions from each project can also be assigned individually.





Row	Movie	Run Time (minutes)	Budget ($M)	USGross ($)	Genre	Dramas	var7
1	White Noise	101	30	56094360	Horror	0	
2	Coach Carter	136	45	67264877	Drama	1	
3	Elektra	100	65	24409722	Action	0	
4	Racing Stripes	110	30	49772522	Comedy	0	
5	Assault on Precinct 13	109	30	20040895	Action	0	
6	Are We There Yet?	94	20	82674398	Comedy	0	
7	Alone in the Dark	96	20	5178569	Horror	0	
8	Indigo	105	25	51100486	Drama	1	
9	Boogeyman	88	20	46752382	Horror	0	
10	The Wedding Date	90	40	31726995	Comedy	0	
11	Hitch	113	55	1.7778426e8	Comedy	0	
12	Pooh's Heffalump Movie	68	20	18081626	Adventure	0	
13	Because of Winn-Dixie	106	15	32647042	Comedy	0	
14	Constantine	122	75	75976178	Action	0	
15	Son of the Mask	86	100	17018422	Comedy	0	
16	Cursed	105	35	19294901	Horror	0	
17	Man of the House	100	50	19699706	Comedy	0	
18	Be Cool	150	75	55849401	Comedy	0	
19	The Jacket	103	28.5	6301131	Thriller	0	
20	The Pacifier	97	56	1.1300688e8	Comedy	0	
21	Hostage	108	75	34636443	Action	0	
22	Robots	89	80	1.282000168	Adventure	0	
23	The Upside of Anger ◆	117	12	18761993	Drama	1	
24	Ice Princess	99	25	24381334	Comedy	0	
25	The Ring Two	111	50	75941727	Horror	0	
26	Miss Congeniality 2: Armed a	115	60	48478006	Comedy	0	

UPDATED! Real-World Data

Statistical concepts are applied to everyday life through the extensive current, real-world data examples and exercises provided throughout the text.

Resources for Success

Instructor Resources

Instructor's Solutions Manual (Download Only)

This manual contains detailed solutions to all of the exercises. These files can be downloaded from within MyLab Statistics or from **www.pearsonglobaleditions.com**.

TestGen

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's online catalog, **www.pearsonglobaleditions.com**. The questions are also assignable in MyLab Statistics.

PowerPoint Lecture Slides

PowerPoint Lecture Slides provide an overview of each chapter, stressing important definitions and offering additional examples. They can be downloaded from MyLab Statistics or from **www.pearsonglobaleditions.com**.

Learning Catalytics

Now included in all MyLab Statistics courses, this student response tool uses students' smartphones, tablets, or laptops to engage them in more interactive tasks and thinking during lecture. Learning Catalytics™ fosters student engagement and peer-to-peer learning with real-time analytics. Access pre-built exercises created specifically for statistics.

Question Libraries

In addition to Statcrunch Projects, MyLab Statistics also includes a Getting Ready for Statistics library that contains more than 450 exercises on prerequisite topics and a Conceptual Question Library with 1000 questions that assess conceptual understanding.

Accessibility

Pearson works continuously to ensure our products are as accessible as possible to all students. We are working toward achieving WCAG 2.0 Level AA and Section 508 standards, as expressed in the Pearson Guidelines for Accessible Educational Web Media, **www.pearson.com/mylab/statistics/accessibility**.

Student Resources

Video Resources

Step-by-Step Example videos guide students through the process of analyzing a problem using the "Think, Show, and Tell" strategy from the textbook. StatTalk Videos, hosted by fun-loving statistician Andrew Vickers, demonstrates important statistical concepts through interesting stories and real-life events. StatTalk videos come with accompanying MyLab assessment questions.

StatCrunch

StatCrunch® is powerful web-based statistical software that allows users to collect, crunch, and communicate with data. The vibrant online community offers tens of thousands of shared datasets for students and instructors to analyze, in addition to all of the datasets in the text or online homework. StatCrunch is integrated directly into MyLab Statistics.

Datasets Available Online

Datasets can be found at the book's webpage on **pearsonglobaleditions.com**. Datasets can be easily transferred to any statistics program.

Statistical Software Support

Instructors and students can copy datasets from the text and MyLab exercises directly into software such as StatCrunch, Data desk, or Excel®. Students can also access instructional support tools including tutorial videos, Study Cards, and manuals for a variety of statistical software programs including StatCrunch, Excel, Minitab®, JMP®, R, SPSS, and TI 83/84 calculators.

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, Portfolio Management Deirdre Lynch was central to the genesis, development, and realization of this project from day one. Our Portfolio Manager, Patrick Barbera, has been invaluable in his support of this edition. Tara Corpuz, Content Producer, kept the cogs from getting into the wheels, where they often wanted to wander. Product Marketing Manager Emily Ockay and Field Marketing Manager Andrew Noble made sure the word got out. Morgan Danna, Portfolio Management Assistant, and Shannon McCormack, Marketing Assistant, were essential in managing all of the behind-the-scenes work. Associate Producer Shannon Bushee put together a top-notch media package for this book. Bob Carroll, Manager of Content Development, for working to ensure digital content meets the mark." Senior Project Manager Chere Bemelmans of the Content Services Centre led us expertly through every stage of production. Manufacturing Buyer Carol Melville, LSC Communications, worked miracles to get this book in your hands.

We would like to draw attention to three people who provided substantial help and support on this edition. First, we would like to acknowledge Nick Horton of Amherst College for his in-depth discussions, guidance, and insights. Nick was invaluable in helping us find the balance between the poles of Normal-based inference and resampling methods. Second, we would like to thank Corey Andreasen of Qatar Academy Doha, Doha, Qatar, and Jared Derksen of Rancho Cucamonga High School for their help with updating the exercises, answers, and datasets.

We'd also like to thank our accuracy checker, 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

Acknowledgments for the Global Edition

Pearson would like to acknowledge and thank the following for the Global Edition:

Contributors

Lynn Farah
ESCP Europe

Dharini Pathmanathan
University of Malaya

Monique Sciortino
University of Malta

David Suda
Lancaster University

Reviewers

Ruben Garcia Berasategui
Jakarta International College

Margarita Medina
University of Canberra

Kiran Paul
Institute of Bioinformatics and Applied Biotechnology

Accounting

Advertising

Agriculture

Banking

Business (General)

Company Names

Consumers

Demographics

Distribution and Operations Management

E-Commerce

Economics

Education

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

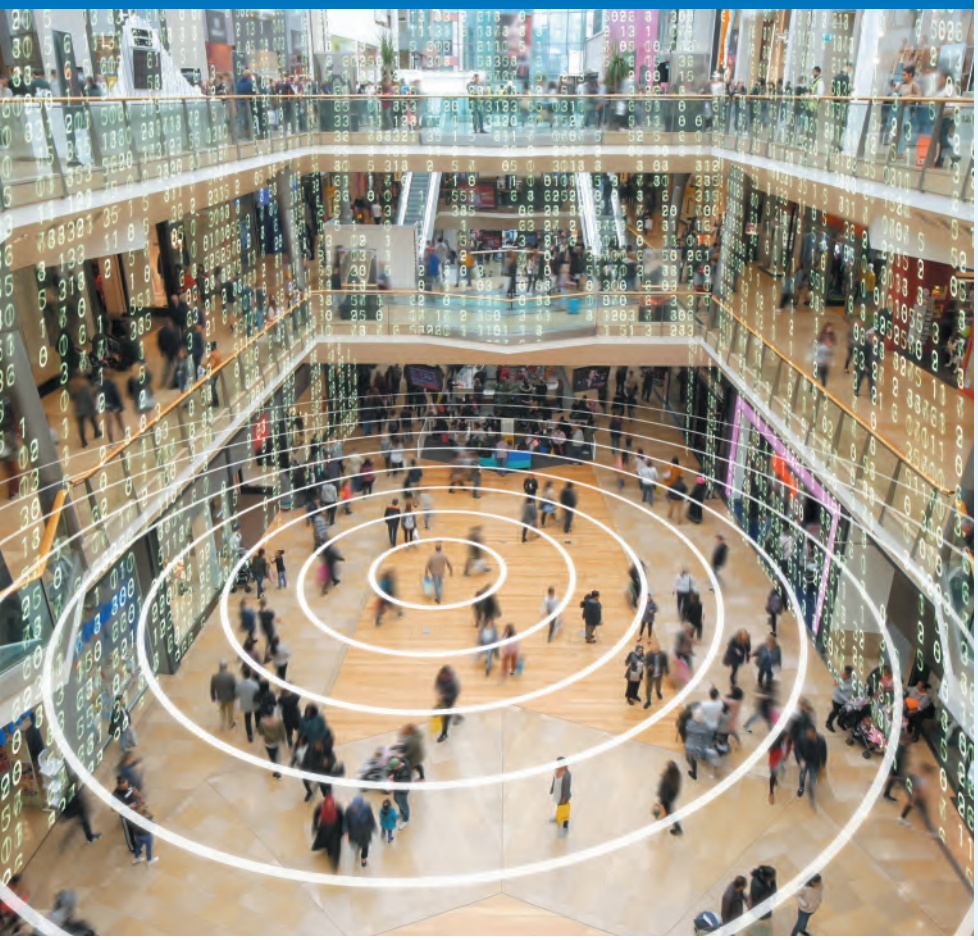
Service Industries and Social Issues

Sports

Surveys and Opinion Polls

Technology

Transportation



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

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

—Lewis Carroll,
Alice's Adventures in Wonderland

1.1 What Is Statistics?

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

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

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

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

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

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

Q: What are statistics?
A: Statistics (plural) are particular calculations made from data.

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

Q: OK, OK, so what are data?
A: Data are values along with their context.

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

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

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

Consider the following:

- If you have a Facebook account, you have probably noticed that the ads you see online tend to match your interests and activities. Coincidence? Hardly. According to *The Wall Street Journal* (10/18/2010),[2] much of your personal information has probably been sold to marketing or tracking companies. Why would Facebook give you a free account and let you upload as much as you want to its site? Because your data are valuable! Using your Facebook profile, a company might build a profile of your interests and activities: what movies and sports you like; your age, sex, education level, and hobbies; where you live; and, of course, who your friends are and what *they* like. From Facebook's point of view, your data are a potential gold mine. Gold ore in the ground is neither very useful nor pretty. But with skill, it can be turned into something both beautiful and valuable. What we're going to talk about 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. Trillions of 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 text, 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.

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 2017, the company's sales reached almost $178 billion (more than 30% 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,

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

[3] informatemi.com/blog/?p=133

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

and even audio and video. If you can sense it, you can measure it. Data are now being collected automatically at such a rate that IBM estimates that "90% of the data in the world today has been created in the last two years alone."[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.

Who and What

In general, the rows of a data table correspond to individual **cases** about *whom* (or about which, if they're not people) we record some characteristics. Cases go by different names, depending on the situation.

- Individuals who answer a survey are called **respondents**.
- People on whom we experiment are **subjects** or (in an attempt to acknowledge the importance of their role in the experiment) **participants**.

[5] www-01.ibm.com/software/data/bigdata/what-is-big-data.html

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."

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

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

Often the cases are a **sample** from some larger **population** that we'd like to understand. Amazon doesn't care about just these customers; it wants to understand the buying patterns of *all* its customers, and, generalizing further, it wants to know how to attract other Internet users who may not have made a purchase from Amazon's site. To be able to generalize from the sample of cases to the larger population, we'll want the sample to be *representative* of that population—a kind of snapshot image of the larger world.

We must know *who* and *what* to analyze data. Without knowing these two, we don't have enough information to start. Of course, we'd always like to know more. The more we know about the data, the more we'll understand about the world. If possible, we'd like to know the *when* and *where* of data as well. Values recorded in 1803 may mean something different than similar values recorded last year. Values measured in Tanzania may differ in meaning from similar measurements made in Mexico. And knowing *why* the data were collected can tell us much about its reliability and quality.

How the Data Are Collected

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

In 2015, *Consumer Reports* published an evaluation of 126 computer 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.