# STATS

## MODELING THE WORLD

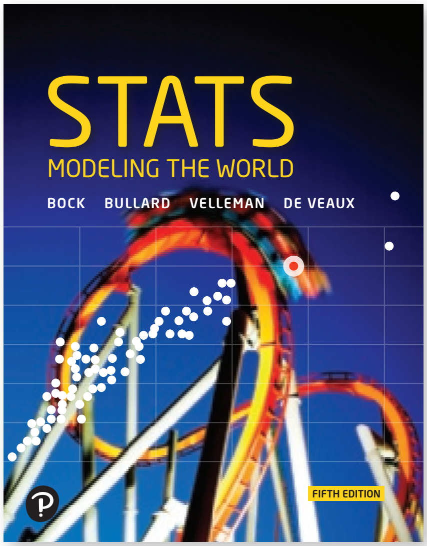
BOCK BULLARD VELLEMAN DE VEAUX

FIFTH EDITION

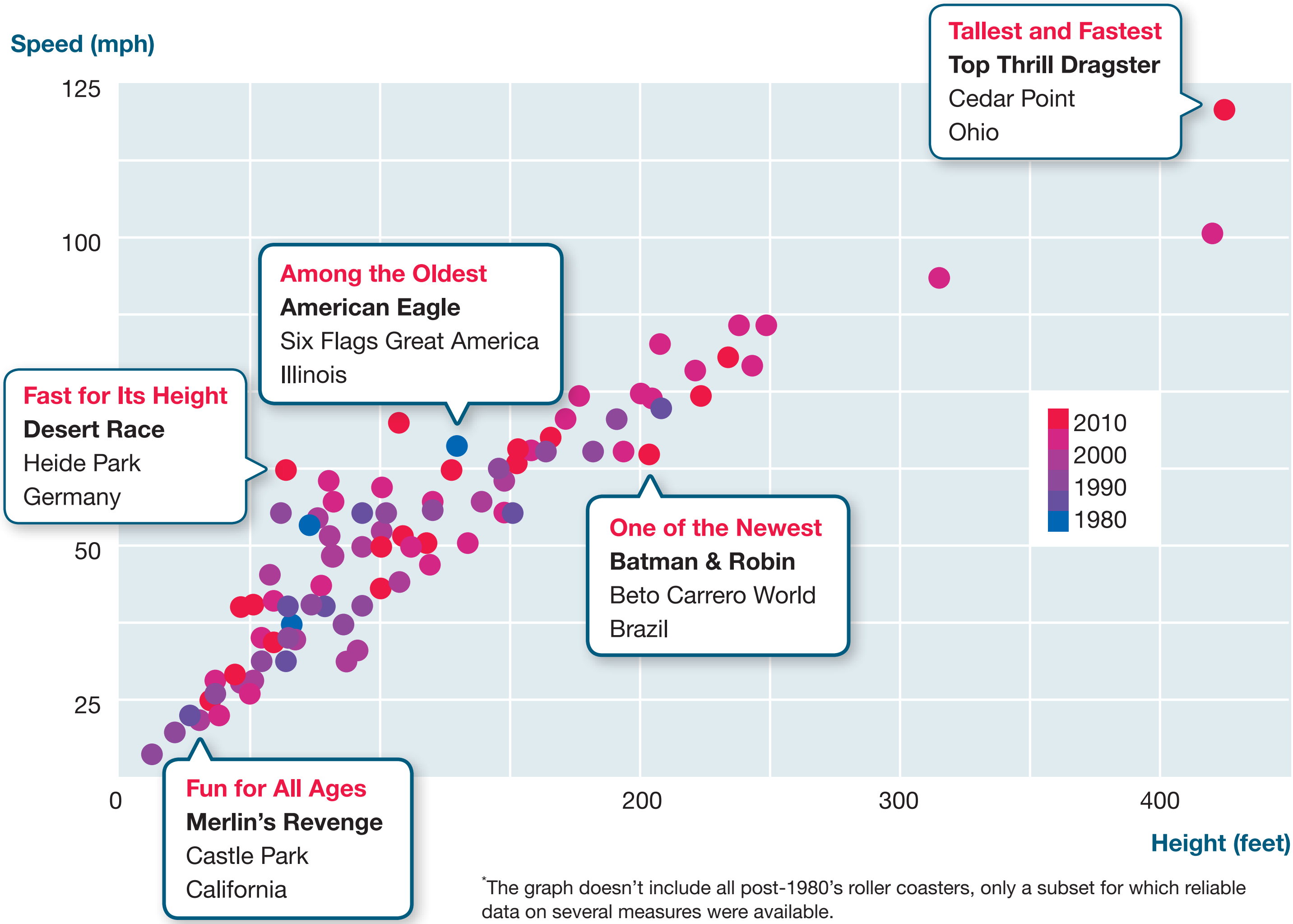
# About the Cover

When the authors and editors were discussing what the cover of this book should look like, we agreed that we wanted it to convey the sense of excitement and fun that we feel when we learn new things from data. We hope you'll share some of our excitement too—Statistics is *fun*! But we were also pretty sure we wanted a real graph showing real data on the cover that would invite readers to explore further. The cover you see is our happy mix of whimsical fun (roller coaster!) and more serious fun (data!). We've reproduced here the graph that's on the cover, but we've added more details.

Each point in the graph represents a roller coaster that opened in 1980 or later.* The horizontal location of a point indicates how tall it is, in feet, and its vertical location indicates the coaster's top speed, in miles per hour (mph). Because colors can provide more information without distracting, we took advantage of color to indicate when each coaster opened: bluer points for older coasters and redder points for newer ones. We also chose five of the coasters and identified them by name.



**What story or stories does the graph tell? What questions does it make you want to ask the data?**



**Tallest and Fastest**
**Top Thrill Dragster**
Cedar Point
Ohio

**Among the Oldest**
**American Eagle**
Six Flags Great America
Illinois

**Fast for Its Height**
**Desert Race**
Heide Park
Germany

**One of the Newest**
**Batman & Robin**
Beto Carrero World
Brazil

**Fun for All Ages**
**Merlin's Revenge**
Castle Park
California

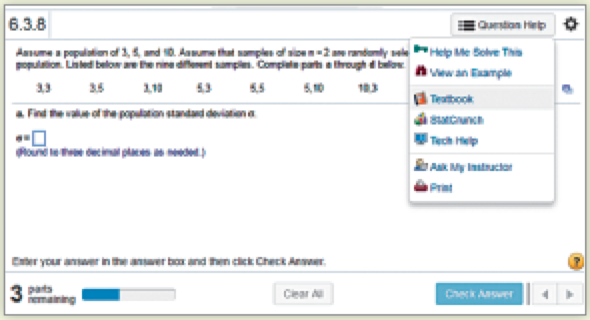
*The graph doesn't include all post-1980's roller coasters, only a subset for which reliable data on several measures were available.

# Get the Most Out of MyLab Statistics

MyLab™ Statistics, Pearson's online homework, tutorial, and assessment program, creates personalized experiences for students and provides powerful tools for instructors. With a wealth of tested and proven resources, each course can be tailored to fit your specific needs. Talk to your Pearson representative about ways to integrate MyLab Statistics into your course for the best results.

## Learning in Any Environment

Because classroom formats and student needs continually change and evolve, MyLab Statistics has built-in flexibility to accommodate various course designs and formats. With a new, streamlined, mobile-friendly design, students and instructors can access courses from most mobile devices to work on exercises and review completed assignments.



6.3.8

Assume a population of 3, 5, and 10. Assume that samples of size n = 2 are randomly sele
population. Listed below are the nine different samples. Complete parts a through d below.

3,3 3,5 3,10 5,3 5,5 5,10 10,3

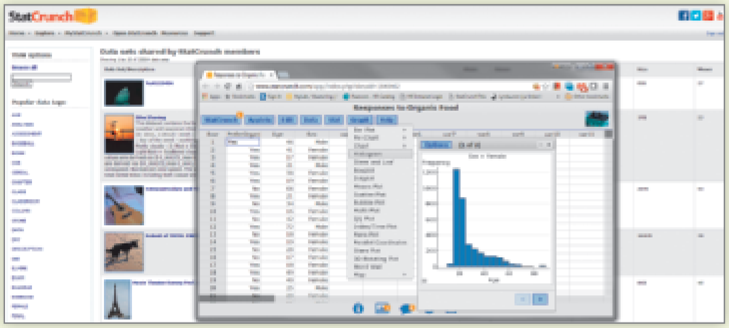
a. Find the value of the population standard deviation σ.

σ = □
(Round to three decimal places as needed.)

Enter your answer in the answer box and then click Check Answer.

## Data Analysis and Interpretation in StatCrunch

Integrated directly into MyLab Statistics, StatCrunch® is powerful web-based statistical software that allows users to perform complex analyses, share data sets, and generate compelling reports of their data.



- **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. A Featured Data page houses the best data sets, making it easy for instructors to use current data in their course. Data sets from the text and from online homework exercises can also be accessed and analyzed in StatCrunch. 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.

**Visit pearson.com/mylab/statistics and click Training & Support to make sure you're getting the most out of MyLab Statistics.**

FIFTH EDITION

# STATS

## MODELING THE WORLD

**DAVID E. BOCK**
Ithaca High School (Retired)

**FLOYD BULLARD**
North Carolina School of Science and Mathematics

**PAUL F. VELLEMAN**
Cornell University

**RICHARD D. DE VEAUX**
Williams College

Pearson

**Director, Portfolio Management:** *Deirdre Lynch*
**Courseware Portfolio Manager:** *Patrick Barbera*
**Courseware Portfolio Assistant:** *Morgan Danna*
**Content Producer:** *Lauren Morse*
**Managing Producer:** *Scott Disanno*
**Producer:** *Nicholas Sweeny*
**Manager, Courseware QA:** *Mary Durnwald*
**Manager, Content Development:** *Robert Caroll*
**Product Marketing Manager:** *Emily Ockay*
**Field Marketing Manager:** *Andrew Noble*
**Marketing Assistants:** *Shannon McCormack, Erin Rush*
**Senior Author Support/Technology Specialist:** *Joe Vetere*
**Manager, Rights and Permissions:** *Gina Cheselka*
**Manufacturing Buyer:** *Carol Melville, LSC Communications*
**Text Design, Production Coordination, Composition, and Illustrations:**
*Cenveo® Publisher Services*
**Cover Image:** *P Wei/E+/Getty Images*

**Library of Congress Cataloging-in-Publication Data**

Names: Bock, David E., author. | Bullard, Floyd, author. | Velleman, Paul F.,
1949—author. | De Veaux, Richard D., author.
Title: Stats : modeling the world / David E. Bock (Ithaca High School (retired)),
Floyd Bullard (North Carolina School of Science and Mathematics), Paul F.
Velleman (Cornell University), Richard D. De Veaux (Williams College).
Other titles: Statistics
Description: 5e [fifth edition]. | Boston, MA : Pearson Education, Inc., [2019]
Identifiers: LCCN 2017045974 | ISBN 9780134685762 (hardcover : alk. paper)
Subjects: LCSH: Graphic calculators—Textbooks. | Mathematical statistics—Textbooks
Classification: LCC QA276.12 .B628 2019 | DDC 519.5—dc23
LC record available at **https://lccn.loc.gov/2017045974**

1 18

Pearson

www.pearson.com

ISBN 13: 978-0-13-468576-2 (Student Edition)
ISBN 10: 0-13-468576-8 (Student Edition)

*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*

*For my partner Rick, who is always supportive and caring and*
*loving, and to whom I am so grateful for everything*

*—Floyd*

*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*

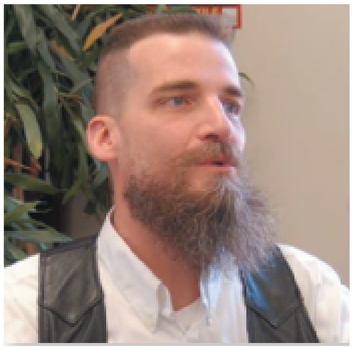
*To Sylvia, who has helped me in more ways than she'll ever know,*
*and to Nicholas, Scyrine, Frederick, and Alexandra,*
*who make me so proud in everything that they are and do*

*—Dick*

# MEET THE AUTHORS

**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.



**Floyd Bullard** first taught high school math as a Peace Corps volunteer in Benin, West Africa, when he was 23 years old. Today he teaches at the North Carolina School of Science and Mathematics in Durham, North Carolina, where he has been since 1999. Floyd has served on the AP Statistics test development committee and presents regularly at workshops and conferences for Statistics teachers.

Floyd's academic degrees are from the Johns Hopkins University (B.S., Applied Mathematics, 1991), the University of North Carolina at Chapel Hill (M.S., Statistics, 1997), and Duke University (Ph.D., Statistics, 2009). He likes to do crossword puzzles and play the piano (not at the same time!).

**Paul F. Velleman** has an international reputation for innovative Statistics education. He is the author and designer of the multimedia Statistics program *ActivStats*, for which he was awarded the EDUCOM Medal for innovative uses of computers in teaching Statistics and the ICTCM Award for Innovation in Using Technology in College Mathematics. He also developed the award-winning Statistics program *Data Desk* and the Internet site Data and Story Library (DASL) (lib.stat.cmu.edu/DASL/), which provides data sets for teaching Statistics. Paul's understanding of using and teaching with technology informs much of this book's approach.

Paul taught Statistics at Cornell University in the Department of Statistical Sciences, for which he was awarded the MacIntyre Prize for Exemplary Teaching. 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.

**Richard D. De Veaux** is an internationally known educator and consultant. He has taught at the Wharton School and the Princeton University School of Engineering, where he won a "Lifetime Award for Dedication and Excellence in Teaching." Since 1994, he has taught at Williams College. 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 is also well known in industry, where for more than 25 years he has consulted for such Fortune 500 companies as American Express, Hewlett-Packard, Alcoa, DuPont, Pillsbury, General Electric, and Chemical Bank. Because he consulted with Mickey Hart on his book *Planet Drum*, he has also sometimes been called the "Official Statistician for the Grateful Dead." His real-world experiences and anecdotes illustrate many of this book's chapters.

Dick holds degrees from Princeton University in Civil Engineering (B.S.E.) and Mathematics (A.B.) and from Stanford University in Dance Education (M.A.) and Statistics (Ph.D.), where he studied dance with Inga Weiss and Statistics with Persi Diaconis.

# TABLE OF CONTENTS

*Optional chapter

## About the Book

Yes, a preface is supposed to be "about this book"—and we'll get there—but first we want to talk about the bigger picture: the ongoing growth of interest in Statistics. These days it seems Statistics is everywhere, from Major League Baseball's innovative StatsCast analytics to the challenges of predicting election outcomes to *Wall Street Journal* and *New York Times* articles touting the explosion of job opportunities for graduates with degrees in Statistics. Public awareness of the widespread applicability, power, and importance of statistical analysis has never been higher. Each year, more students sign up for Stats courses and discover what drew us to this field: it's interesting, stimulating, and even fun. Statistics helps students develop key tools and critical thinking skills needed to become well-informed consumers, parents, and citizens. We think Statistics isn't as much a math course as a civics course, and we're delighted that our books can play a role in preparing a generation for life in the Information Age.

## New to the Fifth Edition

This new edition of *Stats: Modeling the World* extends the series of innovations pioneered in our books, teaching Statistics and statistical thinking as it is practiced today. We've made some important revisions and additions, each with the goal of making it even easier for students to put the concepts of Statistics together into a coherent whole.

- ***More extensive, and more integrated, use of simulations.*** Previous editions all included simulations, but in the fifth edition we've incorporated even more of them, and they're now integrated more fluidly with the text. There's hardly a chapter that doesn't use simulations to motivate a new topic, or to illustrate a concept, or to assist in analyzing data when traditional methods requiring strong assumptions or cumbersome computations are insufficient.

- ***Applets.*** Margin pointers alert students to an innovative set of applets allowing them to explore important concepts and develop deeper understanding of key ideas. Among these: What does standard deviation mean? How might outliers affect our analyses? What does a correlation reveal about a relationship? How does linear regression work? How large should a sample be? What does the all-important Central Limit Theorem say? What does "95% confident" mean? How does hypothesis testing work, and what is a P-value? What are power and Type I and II errors, and how are they interrelated? The applets are found on the book's resource site (www.pearsonhighered.com/bock).

- ***Updated examples, exercises, and data.*** We've updated our innovative *Think/Show/Tell Step-by-Step* examples with new contexts and data. We've added hundreds of new exercises and updated continuing exercises with the most recent data. Whenever possible, we've provided those data on the book's resource site (www.pearsonhighered.com/bock). Most of the examples and exercises are based on recent news stories, research articles, and other real-world sources. We've listed many of those sources so students can explore them further.

# Our Goal: Read This Book!

The best text in the world is of little value if students don't read it. Starting with the first edition, our goal has been to create a book that students would willingly read, easily learn from, and even like. We've been thrilled with the glowing feedback we've received from instructors and students using the first four editions of *Stats: Modeling the World*. Our conversational style, our interesting anecdotes and examples, and even our humor[1] engage students' interest as they learn statistical thinking. We hear from grateful instructors that their students actually do read this book (sometimes even voluntarily reading ahead of the assignments). And we hear from (often amazed) students that they actually enjoyed their textbook.

Here are some of the ways we have made *Stats: Modeling the World*, Fifth Edition, engaging:

- ***Readability.*** You'll see immediately that this book doesn't read like other Statistics texts. The style is both colloquial and informative, enticing students to actually read the book to see what it says.
- ***Informality.*** Our informal style doesn't mean that the subject matter is covered superficially. Not only have we tried to be precise, but 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, making it easier for both instructors and students to focus on one topic at a time.
- ***Consistency.*** We've worked hard to demonstrate how to do Statistics well. From the very start and throughout the book we model the importance of plotting data, of checking assumptions and conditions, and of writing conclusions that are clear, complete, concise, and in context.
- ***The need to read.*** Because the important concepts, definitions, and sample solutions aren't set in boxes, students won't find it easy to just to skim this book. We intend that it be read, so we've tried to make the experience enjoyable.

# Continuing Features

Along with the improvements we've made, you'll still find the many engaging, innovative, and pedagogically effective features responsible for the success of our earlier editions.

- ***Chapter 1 (and beyond).*** Chapter 1 gets down to business immediately, looking at data. And throughout the book chapters lead with new up-to-the-minute motivating examples and follow through with analyses of the data, and real-world examples provide a basis for sample problems and exercises.
- ***Think, Show, Tell.*** The worked examples repeat the mantra of *Think, Show*, and *Tell* in every chapter. They emphasize the importance of thinking about a Statistics question (What do we know? What do we hope to learn? Are the assumptions and conditions satisfied?) and reporting our findings (the *Tell* step). The *Show* step contains the mechanics of calculating results and conveys our belief that it is only one part of the process.
- ***Step-by-Step examples*** guide students through the process of analyzing a problem by showing the general explanation on the left and the worked-out solution on the right. The result: better understanding of the concept, not just number crunching.

[1] And, yes, those footnotes!

- ***For Example.*** In every chapter, an interconnected series of *For Example* elements present a continuing discussion, recapping a story and moving it forward to illustrate how to apply each new concept or skill.

- ***Just Checking.*** At key points in each chapter, we ask students to pause and think with questions designed to be a quick check that they understand the material they've just read. Answers are at the end of the exercise sets in each chapter so students can easily check themselves.

- ***Updated TI Tips.*** Each chapter's easy-to-read "TI Tips" now show students how to use TI-84 Plus CE Statistics functions with the StatWizard operating system. (Help using a TI-Nspire appears in Appendix B, and help with a TI-89 is on the book's resource site www.pearsonhighered.com/bock.) As we strive for a sound understanding of formulas and methods, we want students to use technology for actual calculations. We do emphasize that calculators are just for "Show"—they cannot Think about what to do or Tell what it all means.

- ***Math Boxes.*** In many chapters we present the mathematical underpinnings of the statistical methods and concepts. By setting these proofs, derivations, and justifications apart from the narrative, we allow students to continue to follow the logical development of the topic at hand, yet also explore the underlying mathematics for greater depth.

- ***TI-Nspire Activities.*** Margin pointers identify demonstrations and investigations for TI-Nspire handhelds to enhance each chapter. They're found at the book's resource site (www.pearsonhighered.com/bock).

- ***What Can Go Wrong?*** Each chapter still contains our innovative *What Can Go Wrong?* sections that highlight the most common errors people make and the misconceptions they have about Statistics. Our goals are to help students avoid these pitfalls and to arm them with the tools to detect statistical errors and to debunk misuses of statistics, whether intentional or not.

- ***What Have We Learned?*** Chapter-ending study guides help students review key concepts and terms.

- ***Exercises.*** We've maintained the pairing of examples so that each odd-numbered exercise (with an answer in the back of the book) is followed by an even-numbered exercise illustrating the same concept. Exercises are ordered by approximate level of complexity.

- ***Practice Exams.*** At the end of each of the book's seven parts you'll find a practice exam, consisting of both multiple choice and free response questions. These cumulative exams encourage students to keep important concepts and skills in mind throughout the course while helping them synthesize their understanding as they build connections among the various topics.

- ***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 Alerts.*** Clear communication is essential in Statistics, and proper notation is part of the vocabulary students need to learn. We've found that it helps to call attention to the letters and symbols statisticians use to mean very specific things.

- ***On the Computer.*** Because real-world data analysis is done on computers, at the end of each chapter we summarize what students can find in most Statistics software, usually with an annotated example.

# Our Approach

We've been guided in the choice of topics and emphasis on clear communication by the requirements of the Advanced Placement Statistics course. In our order of presentation, we have tried to ensure that each new topic fits logically into the growing structure of understanding that we hope students will build.

## GAISE Guidelines

We have worked to provide materials to help each class, in its own way, follow the guidelines of the GAISE (Guidelines for Assessment and Instruction in Statistics Education) project sponsored by the American Statistical Association. That report urges that Statistics education should

1. emphasize statistical literacy and develop statistical thinking,
2. use real data,
3. stress conceptual understanding rather than mere knowledge of procedures,
4. foster active learning,
5. use technology for developing concepts and analyzing data, and
6. make assessment a part of the learning process.

## Mathematics

Mathematics traditionally appears in Statistics texts in several roles:

1. It can provide a concise, clear statement of important concepts.
2. It can embody proofs of fundamental results.
3. It can describe calculations to be performed with data.

Of these, we emphasize the first. Mathematics can make discussions of Statistics concepts, probability, and inference clear and concise. We have tried to be sensitive to those who are discouraged by equations by also providing verbal descriptions and numerical examples.

This book is not concerned with proving theorems about Statistics. Some of these theorems are quite interesting, and many are important. Often, though, their proofs are not enlightening to Introductory Statistics students and can distract the audience from the concepts we want them to understand. However, we have not shied away from the mathematics where we believed that it helped clarify without intimidating. You will find some important proofs, derivations, and justifications in the Math Boxes that accompany the development of many topics.

Nor do we concentrate on calculations. Although statistics calculations are generally straightforward, they are also usually tedious. And, more to the point, they are often unnecessary. Today, virtually all statistics are calculated with technology, so there is little need for students to work by hand. The equations we use have been selected for their focus on understanding concepts and methods.

# Technology and Data

To experience the real world of Statistics, it's best to explore real data sets using modern technology. This fact permeates *Stats: Modeling the World*, Fifth Edition, where we use real data for the book's examples and exercises. Technology lets us focus on teaching statistical thinking rather than getting bogged down in calculations. The questions that motivate each of our hundreds of examples are not "How do you find the answer?" but "How do you think about the answer?"

**Technology.** We assume that students are using some form of technology in this Statistics course. That could include a graphing calculator along with a Statistics package or spreadsheet. Rather than adopt any particular software, we discuss generic computer output. "TI-Tips"—included in most chapters—show students how to use Statistics features of the TI-84 Plus series. In Appendix B, we offer general guidance (by chapter) to help students get started on common software platforms (StatCrunch, Excel, MINITAB, Data Desk, and JMP) and a TI-Nspire. The book's resource site (www.pearsonhighered.com/bock) includes additional guidance for students using a TI-89. Students will also find on the site applets that let them explore key concepts.

**Data.** Because we use technology for computing, we don't limit ourselves to small, artificial data sets. In addition to including some small data sets, we have built examples and exercises on real data with a moderate number of cases—usually more than you would want to enter by hand into a program or calculator. These data are included on the book's resource site, www.pearsonhighered.com/bock.

# MyLab Statistics Online Course (access code required)

Used by nearly one million students a year, MyLab Statistics is the world's leading online program for teaching and learning Statistics. MyLab Statistics delivers assessment, tutorials, and multimedia resources that provide engaging and personalized experiences for each student, so learning can happen in any environment.

## Personalized Learning

Not every student learns the same way or at the same rate. Personalized learning in the MyLab gives instructors the flexibility to incorporate the approach that best suits the needs of their course and students.

- Based on their performance on a quiz or test, students can be given **personalized homework** to allow them to focus on just the topics they have not yet mastered.
- With **Companion Study Plan Assignments**, instructors can assign the Study Plan as a prerequisite to a test or quiz, guiding students through the concepts they need to master.

## Preparedness

Preparedness is one of the biggest challenges in Statistics courses. Pearson offers a variety of content and course options to support students with just-in-time remediation and key-concept review as needed.

- **Redesign-Ready Course Options:** Many new course models have emerged in recent years, as institutions "redesign" to help improve retention and results. At Pearson, we're focused on tailoring solutions to support instructors' plans and programs.
- **Getting Ready for Statistics Questions:** This question library contains more than 450 exercises that cover the relevant developmental math topics for a given section. These can be made available to students for extra practice or assigned as a prerequisite to other assignments.

## Conceptual Understanding

Successful students have the ability to apply their statistical ideas and knowledge to new concepts and real-world situations. Providing frequent opportunities for data analysis and interpretation helps students develop the 21st century skills that they need to be successful in the classroom and workplace.

- **Conceptual Question Library:** There are 1,000 questions in the Assignment Manager that require students to apply their statistical understanding.
- **Modern statistics is practiced with technology**, and MyLab Statistics makes learning and using software programs seamless and intuitive. Instructors can copy data sets from the text and MyLab Statistics exercises directly into software such as StatCrunch 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.

## Motivation

Students are motivated to succeed when they're engaged in the learning experience and understand the relevance and power of Statistics.

- **Exercises with Immediate Feedback:** Homework and practice exercises in MyLab Statistics regenerate algorithmically to give students unlimited opportunity for practice and mastery. Instructors can choose from the many exercises available for the author's approach—or even choose additional exercises from other MyLab Statistics courses. Most exercises include learning aids, such as guided solutions, sample problems, extra help at point-of-use, and immediate feedback when students enter incorrect answers.
- Instructors can create, import, and manage online homework assignments, quizzes, and tests—or start with sample assignments—all of which are automatically graded, allowing instructors to spend less time grading and more time teaching.

## Data & Analytics

MyLab Statistics provides resources to help instructors assess and improve student results. A comprehensive Gradebook with enhanced reporting functionality makes it easier for instructors to manage courses efficiently.

- **Reporting Dashboard:** Instructors can view, analyze, and report learning outcomes, gaining the information they need to keep students on track. Available via the Gradebook and fully mobile-ready, the Reporting Dashboard presents student performance data at the class, section, and program levels in an accessible, visual manner. Its fine-grain reports allow instructors and administrators to compare performance across different courses, across individual sections, and within each course.
- **Item Analysis:** Instructors can track classwide understanding of particular exercises in order to refine lectures or adjust the course/department syllabus. Just-in-time teaching has never been easier.

## 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.

# StatCrunch

Integrated directly into MyLab Statistics, StatCrunch is powerful Web-based statistical software that allows users to perform complex analyses, share data sets, and generate compelling reports of their data.

- **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. A Featured Data page houses the best data sets, making it easy for instructors to use current data in their course. Data sets from the text and from online homework exercises can also be accessed and analyzed in StatCrunch. An online survey tool allows users to quickly collect data via Web-based surveys.
- **Crunch.** A full range of numerical and graphical methods allows 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.

StatCrunch is integrated into MyLab Statistics, but it is also available by itself to qualified adopters. StatCrunch is also now available on your smartphone or tablet when you visit www.statcrunch.com from the device's browser. For more information, visit our Web site at www.statcrunch.com, or contact your Pearson representative.

## MathXL Online Course (access code required)

Part of the world's leading collection of online homework, tutorial, and assessment products, MathXL® delivers assessment and tutorial resources that provide engaging and personalized experiences for each student. Each course is developed to accompany Pearson's best-selling content, authored by thought leaders across the math curriculum, and can be easily customized to fit any course format.

**With MathXL, instructors can:**

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

**With MathXL, students can:**

- Take chapter tests in MathXL and receive personalized study plans and/or personalized homework assignments based on their test results.
- Use the study plan and/or the homework to link directly to tutorial exercises for the objectives they need to study.
- Access supplemental animations and video clips directly from selected exercises.

MathXL is available to qualified adopters. For more information, visit our Web site at www.pearson.com/MathXL, or contact your Pearson representative.

**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 12-month access to Minitab and Minitab Express ensures 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 a stand-alone).

**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.

**XLSTAT™** is an Excel add-in that enhances the analytical capabilities of Excel. XLSTAT is used by leading businesses and universities around the world. It is available to bundle with this text. For more information, go to www.pearsonhighered.com/xlstat/. ISBN-13: 978-0-321-75932-0; ISBN-10: 0-321-75932-X.

Pearson
MyLab

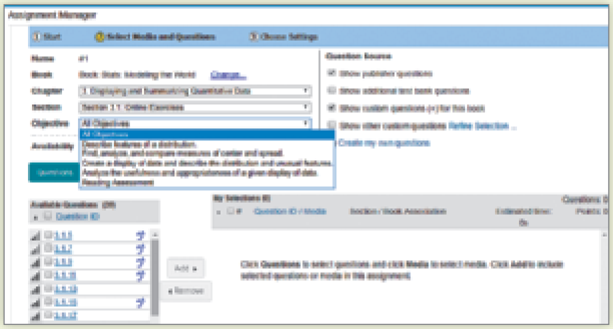
# Resources for Success

## MyLab Statistics Online Course for *Stats: Modeling the World,* 5e, by David E. Bock, Floyd Bullard, Paul F. Velleman, and Richard D. De Veaux (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 is tightly integrated throughout the accompanying MyLab Statistics course, making learning the material as seamless as possible.

### Expanded Objective-Based MathXL Coverage

MathXL® is newly mapped to improve student learning outcomes. Homework reinforces and supports students' understanding of key Statistics topics.



Assignment Manager

① Start ② Select Media and Questions ③ Choose Settings

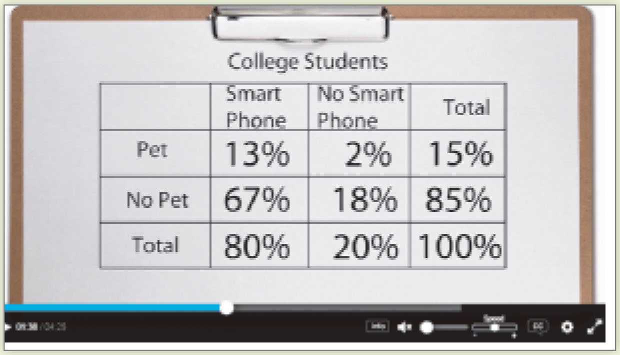
Name #1
Book Book: Stats: Modeling the World Change...
Chapter 1. Displaying and Summarizing Quantitative Data
Section Section 1.1: Online Exercises
Objective All Objectives
All Objectives
Describe features of a distribution.
Find, analyze, and compare measures of center and spread.
Create a display of data and describe the distribution and unusual features.
Analyze the usefulness and appropriateness of a given display of data.
Reading Assessment
Availability

Question Source
☑ Show publisher questions
□ Show additional test bank questions
☑ Show custom questions (+) for this book
□ Show other custom questions Refine Selection ...
Create my own questions

Available Questions (29)
□ Question ID
□ 1.1.1
□ 1.1.2
□ 1.1.3
□ 1.1.11
□ 1.1.13
□ 1.1.15
□ 1.1.17

My Selections (0)
□ # Question ID / Media Section / Book Association Estimated time: 0s Questions: 0 Points: 0

Click Questions to select questions and click Media to select media. Click Add to include selected questions or media in this assignment.



College Students

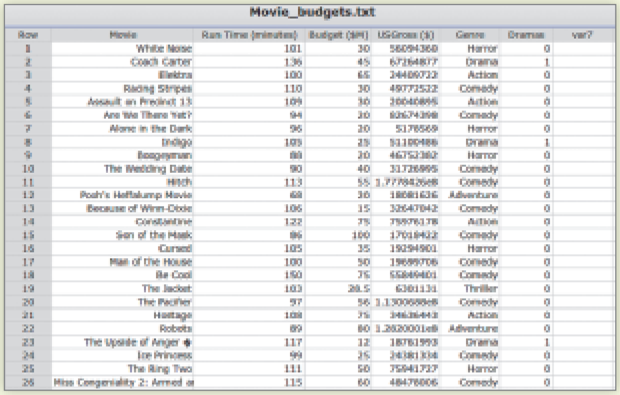
|  | Smart Phone | No Smart Phone | Total |
|---|---|---|---|
| Pet | 13% | 2% | 15% |
| No Pet | 67% | 18% | 85% |
| Total | 80% | 20% | 100% |

### Enhanced Video Program

Updated Step-by-Step Example videos guide students through the process of analyzing a problem using the "Think, Show, and Tell" strategy from the textbook.

### Real-World Data Examples

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



Movie_budgets.txt

| 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 | 1.77784326e8 | Comedy | 0 |  |
| 12 | Pooh's Heffalump Movie | 68 | 30 | 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 | 1.13006988e8 | Comedy | 0 |  |
| 21 | Hostage | 108 | 75 | 34636443 | Action | 0 |  |
| 22 | Robots | 89 | 80 | 1.2820001e8 | 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 |  |

**pearson.com/mylab/statistics**

# Resources for Success

**Instructor's Edition** contains answers to all exercises. (ISBN-13: 978-0-13-468795-7; ISBN-10: 0-13-468795-7)

**Instructor's Solutions Manual (download only)**, by Adam Yankay and William Craine, contains solutions to all the exercises. It can be downloaded from within MyLab Statistics or from Pearson's online catalog, www.pearson.com.

**Online Test Bank and Resource Guide (download only)**, by John Diehl, Jane Viau, and William Craine, 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. It can be downloaded from within MyLab Statistics or from Pearson's online catalog, www.pearson.com.

**TestGen®** (www.pearson.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's online catalog, www.pearson.com. The questions are also assignable in MyLab Statistics.

**PowerPoint Lecture Slides (downloadable)** Classroom presentation slides feature key graphics, concepts, and examples from this text. They can be downloaded from within MyLab Statistics or from Pearson's online catalog, www.pearson.com.

**Learning Catalytics™** is 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 lectures. Learning Catalytics fosters student engagement and peer-to-peer learning with real-time analytics and provides access to pre-built exercises created specifically for Statistics.

**Graphing Calculator Manual (download only)**, by John Diehl (Hinsdale Central High School), includes worked-out examples from the text along with a step-by-step guide on how to use the TI-84 Plus, TI-Nspire, and Casio graphing calculators. It can be downloaded by students and instructors from within MyLab Statistics.

**pearson.com/mylab/statistics**

# ACKNOWLEDGMENTS

Many people have contributed to this book in all five of its editions. This edition would have never seen the light of day without the assistance of the incredible team at Pearson. Our Director of Portfolio Management, Deirdre Lynch, was central to the genesis, development, and realization of the book from day one. Patrick Barbera, Courseware Portfolio Manager, and Emily Ockay, Product Marketing Manager, provided much needed support. Lauren Morse, Content Producer, kept us on task as much as humanly possible. Morgan Danna, Courseware Portfolio Assistant, and Shannon McCormack, Product Marketing Assistant, were essential in managing all of the behind-the-scenes work that needed to be done. Nick Sweeney, Producer, put together a top-notch media package for this book. Carol Melville, Manufacturing Buyer, LSC Communications, worked miracles to get this book in your hands. We'd also like to thank Meredith Douthett who created some new cartoons for this revision. Special thanks go out to Cenveo Publisher Services for the wonderful work they did on this book, and in particular to Chere Bemelmans, Project Manager, for her close attention to detail, for keeping the cogs from getting into the wheels where they often wanted to wander, and for keeping the authors on task and on schedule . . . no easy feat!

We would like to thank Corey Andreasen of Qatar Academy Doha, Doha, Qatar, and Jared Derksen of Rancho Cucamonga High School for their invaluable help with updating the exercises, answers, and data sets. We'd also like to thank our accuracy checkers whose monumental task was to make sure we said what we thought we were saying. They are Susan Blackwell, First Flight High School; Douglas Cashing, St. Bonaventure University; Benjamin Kirk, Ithaca High School; and Paul Lorczak.

We extend our sincere thanks for the suggestions and contributions made by the following reviewers, focus group participants, and class-testers:

John Arko
*Glenbrook South High School, IL*

Kathleen Arthur
*Shaker High School, NY*

Allen Back
*Cornell University, NY*

Beverly Beemer
*Ruben S. Ayala High School, CA*

Judy Bevington
*Santa Maria High School, CA*

Susan Blackwell
*First Flight High School, NC*

Gail Brooks
*McLennan Community College, TX*

Walter Brown
*Brackenridge High School, TX*

Darin Clifft
*Memphis University School, TN*

Bill Craine
*Lansing High School, NY*

Sybil Coley
*Woodward Academy, GA*

Kevin Crowther
*Lake Orion High School, MI*

Caroline DiTullio
*Summit High School, NJ*

Jared Derksen
*Rancho Cucamonga High School, CA*

Sam Erickson
*North High School, WI*

Laura Estersohn
*Scarsdale High School, NY*

Laura Favata
*Niskayuna High School, NY*

David Ferris
*Noblesville High School, IN*

Linda Gann
*Sandra Day O'Connor High School, TX*

Randall Groth
*Illinois State University, IL*

Donnie Hallstone
*Green River Community College, WA*

Howard W. Hand
*St. Marks School of Texas, TX*

Bill Hayes
*Foothill High School, CA*

Miles Hercamp
*New Palestine High School, IN*

Michelle Hipke
*Glen Burnie Senior High School, MD*

Carol Huss
*Independence High School, NC*

Sam Jovell
*Niskayuna High School, NY*

Peter Kaczmar
*Lower Merion High School, PA*

John Kotmel
*Lansing High School, NY*

Beth Lazerick
*St. Andrews School, FL*

Michael Legacy
*Greenhill School, TX*

Guillermo Leon
*Coral Reef High School, FL*

John Lieb
*The Roxbury Latin School, MA*

Mark Littlefield
*Newburyport High School, MA*

Martha Lowther
*The Tatnall School, DE*

John Maceli
*Ithaca College, NY*

Jim Miller
*Alta High School, UT*

Timothy E. Mitchell
*King Philip Regional High School, MA*

Maxine Nesbitt
*Carmel High School, IN*

Elizabeth Ann Przybysz
*Dr. Phillips High School, FL*

Diana Podhrasky
*Hillcrest High School, TX*

Rochelle Robert
*Nassau Community College, NY*

Karl Ronning
*Davis Senior High School, CA*

Bruce Saathoff
*Centennial High School, CA*

Agatha Shaw
*Valencia Community College, FL*

Murray Siegel
*Sam Houston State University, TX*

Chris Sollars
*Alamo Heights High School, TX*

Darren Starnes
*The Webb Schools, CA*

Jane Viau
*The Frederick Douglass Academy, NY*

And we'd be especially remiss if we did not applaud the outstanding team of teachers whose creativity, insight, knowledge, and dedication create the many valuable resources so helpful for Statistics students and their instructors:

Corey Andreasson,
*Qatar Academy Doha*

Anne Carroll
*Kennett High School, PA*

Ruth Carver
*Germantown Academy in Fort Washington, PA*

William B. Craine
*Lansing High School, NY*

Jared Derksen
*Rancho Cucamonga High School, CA*

John Diehl
*Hindsdale Central High School, IL*

Lee Kucera
*Capistrano Valley High School, CA*

John Mahoney
*Benjamin Banneker Academic High School, Washington, D.C.*

Susan Peters
*University of Louisville, KY*

Janice Ricks
*Marple Newtown High School, PA*

Jane Viau
*The Frederick Douglass Academy, NY*

Webster West
*Creator of StatCrunch*

Adam Yankay
*Western Reserve Academy, OH*

*David Bock*
*Floyd Bullard*
*Paul Velleman*
*Richard De Veaux*

1

# Stats Starts Here[1]

> “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

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 a Statistics course may not be your friends’ first choice for a fun elective.

But Statistics *is* fun. That’s probably not what you heard on the street, but it’s true. Statistics is about how to think clearly with data. We’ll talk about data in more detail soon, but for now, think of **data** as any collection of numbers, characters, images, or other items that provide information about something. Whenever there are data and a need for understanding the world, you’ll find Statistics. A little practice thinking statistically is all it takes to start seeing the world more clearly and accurately.

## So, What Is (Are?) Statistics?

How does it feel to live near the dawn of the Information Age? More data are collected today (a lot of it on *us*) than ever before in human history. Consider these examples of how data are used today:

- 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

[1] We could have called this chapter “Introduction,” but nobody reads the introduction, and we wanted you to read this. We feel safe admitting this here, in the footnote, because nobody reads footnotes either.

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

**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.

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.

- Like many other retailers, Target stores create customer profiles by collecting data about purchases using credit cards. Patterns the company discovers across similar customer profiles enable it to send you advertising and coupons that promote items you might be particularly interested in purchasing. As valuable to the company as these marketing insights can be, some may prove startling to individuals. Recently, coupons Target sent to a Minneapolis girl's home revealed she was pregnant before her father knew![3]
- How dangerous is texting while driving? Researchers at the University of Utah tested drivers on simulators that could present emergency situations. They compared reaction times of sober drivers, drunk drivers, and texting drivers.[4] The results were striking. The texting drivers actually responded more slowly and were more dangerous than those who were above the legal limit for alcohol.

With this text you'll learn how to design good studies and discern messages in data, should you become a researcher yourself. More important—especially to those who, like most of us, *don't* become researchers—you'll learn to judge with a more skilled and critical eye the conclusions drawn from data by *others*. With so much data everywhere, you need such judgment just to be an informed and responsible citizen.

**ARE YOU A STATISTIC?**

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."

## Statistics in a Word

It can be fun, and sometimes useful, to summarize a discipline in only a few words. So,

Economics is about . . . *Money (and why it is good).*

Psychology: *Why we think what we think (we think).*

Biology: *Life.*

Anthropology: *Who?*

History: *What, where, and when?*

Philosophy: *Why?*

Engineering: *How?*

Accounting: *How much?*

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

Some of the reasons why we act and think differently from one another can be explained—perhaps by our education or our upbringing or how our friends act and think. But some of that variation among us will always remain unexplained.[5] Statistics is largely about trying to find explanations for why data vary while acknowledging that some amount of the variation will always remain a mystery.

[3] www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#1770c3576668

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

[5] And that's a good thing!

# But What *Are* Data?

Amazon.com opened for business in July 1995, billing itself as "Earth's Biggest Bookstore." By 1997, Amazon had a catalog of more than 2.5 million book titles and had sold books to more than 1.5 million customers in 150 countries. In 2016, the company's sales reached $136 billion (more than a 27% increase from the previous year). Amazon has sold a wide variety of merchandise, including a $400,000 necklace, yak cheese from Tibet, and the largest book in the world. How did Amazon become so successful and how can it keep track of so many customers and such a wide variety of products? The answer to both questions is *data.*

But what are data? Think about it for a minute. What exactly *do* we mean by "data"? Do data have to be numbers? The amount of your last purchase in dollars is numerical data. But your name and address in Amazon's database are also data even though they are not numerical. What about your ZIP code? That's a number, but would Amazon care about, say, the *average* ZIP code of its customers?

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

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 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. |
| Will S. | 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 | 567 | 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 | Will S. | Ohio | 0.99 | 567 | 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 information 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.

THE W'S:
WHO
WHAT and in what units
WHEN
WHERE
WHY
HOW

## *Who* and *What*

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

- Individuals who answer a survey are called **respondents**.
- People on whom we experiment are **subjects** or (in an attempt to acknowledge the importance of their role in the experiment) **participants**.
- Animals, plants, websites, and other inanimate subjects are often called **experimental units**.
- Often we simply call cases what they are: for example, *customers, economic quarters*, or *companies*.
- In a database, rows are called **records**—in this example, purchase records. Perhaps the most generic term is *cases*, but in any event the rows represent the *who* of the data.[6]

The characteristics recorded about each individual are called **variables**. These are usually shown as the columns of a data table, and they should have a name that identifies *What* has been measured. The data table of Amazon purchases shows that some of the variables that Amazon collected data for were *Name*, *Price*, and whether the purchase was a *Gift*.

Often, the cases are a **sample** of cases selected from some larger **population** that we'd like to understand. Amazon certainly cares about its customers, but also wants to know how to attract all those other Internet users who may never 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.

### FOR EXAMPLE

### Identifying the *Who*

*Consumer Reports* included 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 population of interest is all tablets currently available. The sample (and the *Who* of the study) is the particular collection of 126 tablets that the consumer organization purchased and studied.

## How the Data Are Collected

*How* the data are collected can make the difference between insight and nonsense. For example, as we'll see later, data that come from a voluntary survey on the Internet are almost always worthless. One primary concern of Statistics is the design of sound methods

[6] If you're going through the *W*'s in your head and your data are measurements of objects, "*Who*" should still remind you to ask, "What objects were measured?"

for collecting data.[7] 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.

THINK

SHOW

TELL

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.

## More About Variables (*What*?)

The Amazon data table displays information about several variables: *Order Number*, *Name*, *State/Country*, *Price*, and so on. These identify *what* we know about each individual. Variables such as these can play different roles, depending on how we plan to use them. While some are merely identifiers, others may be categorical or quantitative. Making that distinction is an important step in our analysis.

### Identifiers

Amazon wants to know who you are when you sign in again and it doesn't want to confuse you with some other customer. So it assigns you a unique identifying number.[8] Amazon also wants to send you the right product, so it assigns a unique Amazon Standard Identification Number (ASIN) to each item it carries. Both of these numbers are useful to Amazon, but they aren't measurements of anything. They're generated by Amazon and *assigned* uniquely to customers and products. Data like these values are called **identifier variables**. Other examples are student ID numbers and Social Security numbers. Identifier variables are typically used to identify single cases, not to look for patterns in collections of data, so they are seldom used in data analysis.

### Categorical Variables

Some variables just tell us what group or category each individual belongs to. Are you male or female? Pierced or not? What color are your eyes? We call variables like these **categorical variables**.[9] 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. Descriptive responses to questions are often categories. For example, the responses to the questions "Who is your cell phone provider?" or "What is your marital status?" yield categorical values. 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.

PRIVACY AND THE INTERNET

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

[7] Coming attractions: to be discussed in Part III. We sense your excitement.

[8] Or sometimes a code containing numerals and letters.

[9] You may also see them called *qualitative* variables.

## 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 bronze knuts.[10]

## Either/Or?

Some variables with numeric values can be treated as either categorical or quantitative depending on what we want to know. 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.

Suppose a course evaluation survey asks, "How valuable do you think this course will be to you?" 1 = Worthless; 2 = Slightly; 3 = Somewhat; 4 = Reasonably; 5 = Invaluable. Is *Educational Value* categorical or quantitative? A teacher might just count the number of students who gave each response for her course, treating *Educational Value* as a categorical variable. Or if 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 4.2 seems more valuable than one that averages 2.1. But is it *twice* as valuable? Does that even mean anything? Variables like this that have a natural order but no units are often called *ordinal variables*. Other examples are college class (freshman, sophomore, junior, or senior) and hurricane level (1, 2, 3, 4, or 5). Ordinal variables can be a little tricky to analyze and for the most part they are not considered in this text.

FOR EXAMPLE

### Identifying the *What* and *Why* of Tablets

**RECAP:** A *Consumer Reports* article about 126 tablet computers lists each tablet's manufacturer, cost, battery life (hours), operating system (iOS/Android/RIM) and overall performance score (0–100).

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

**ANSWER:** The variables are:

- manufacturer (categorical)
- cost (quantitative, $)
- battery life (quantitative, hours)
- operating system (categorical)
- performance score (quantitative, with no units—essentially an ordinal variable)

*Why?* The magazine hopes to provide consumers with information to help them choose a tablet to meet their needs.

[10] Wizarding money. Just seeing who's paying attention.

## JUST CHECKING

In the 2004 Tour de France, Lance Armstrong made history by winning the race for an unprecedented sixth time. In 2005, he became the only 7-time winner and set a new record for the fastest average speed—41.65 kilometers per hour (about 26 mph)—that stands to this day. Then in 2012, Armstrong was banned for life for doping offenses, stripped of all his titles and his records expunged. Here are the first three and last seven lines of the data set. Keep in mind that the 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 Francaise | 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 |
| . . . | | | | | | | | | | |
| 2011 | Cadel Evans | Australia | 34 | BMC | 86.12.22 | 39.79 | 21 | 3430 | 198 | 167 |
| 2012 | Bradley Wiggins | Great Britain | 32 | Sky | 87.34.47 | 39.83 | 20 | 3488 | 198 | 153 |
| 2013 | Cristopher 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 | Cristopher Froome | Great Britain | 30 | Sky | 84.46.14 | 39.64 | 21 | 3660.3 | 198 | 160 |
| 2016 | Cristopher Froome | Great Britain | 31 | Sky | 89.04.48 | 39.62 | 21 | 3529 | 198 | 174 |
| 2017 | Cristopher Froome | Great Britain | 32 | Sky | 86.20.55 | 38.52 | 21 | 3326.5 | 219 | 188 |

### THERE'S A WORLD OF DATA ON THE INTERNET

These days, one of the richest sources of data is the Internet. With a bit of practice, you can learn to find data on almost any subject. Many of the data sets we use in this book were found in this way. The Internet has both advantages and disadvantages as a source of data. Among the advantages are the fact that often you'll be able to find even more current data than those we present. The disadvantage is that references to Internet addresses can "break" as sites evolve, move, and die.

Our solution to these challenges is to offer the best advice we can to help you search for the data, wherever they may be residing. We usually point you to a website. We'll sometimes suggest search terms and offer other guidance.

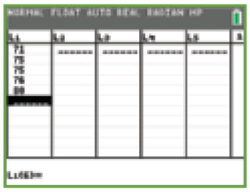
Some words of caution, though: Data found on Internet sites may not be formatted in the best way for use in statistics software. Although you may see a data table in standard form, an attempt to copy the data may leave you with a single column of values. You may have to work in your favorite statistics or spreadsheet program to reformat the data into variables. You will also probably want to remove commas from large numbers and extra symbols such as money indicators ($, ¥, £); few statistics packages can handle these.

## WHAT CAN GO WRONG?

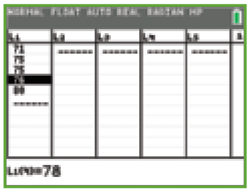
- **Don't label a variable as categorical or quantitative without thinking about the question you want it to answer.** The same variable can sometimes take on different roles.
- **Just because your variable's values are numbers, don't assume that it's quantitative.** Categories are often given numerical labels. Don't let that fool you into thinking they have quantitative meaning. Look at the context.
- **Always be skeptical.** One reason to analyze data is to discover the truth. Even when you are told a context for the data, it may turn out that the truth is a bit (or even a lot) different. Think about *how* and *why* the data were collected.

## TI TIPS

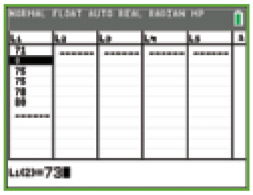
### Working with Data



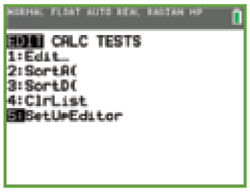
```
NORMAL FLOAT AUTO REAL RADIAN MP
L1    L2    L3    L4    L5
71    ----- ----- ----- -----
75
75
76
80
-----
L1(6)=
```



```
NORMAL FLOAT AUTO REAL RADIAN MP
L1    L2    L3    L4    L5
71    ----- ----- ----- -----
75
75
78
80
-----
L1(4)=78
```



```
NORMAL FLOAT AUTO REAL RADIAN MP
L1    L2    L3    L4    L5
71    ----- ----- ----- -----
73
75
75
78
80
-----
L1(2)=73
```



```
NORMAL FLOAT AUTO REAL RADIAN MP
EDIT CALC TESTS
1:Edit…
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
```

You'll need to be able to enter and edit data in your calculator. Here's how:

**TO ENTER DATA:** Hit the `STAT` button, and choose `EDIT` from the menu. You'll see a set of columns labeled `L1`, `L2`, and so on. Here is where you can enter, change, or delete a set of data.

Let's enter the heights (in inches) of the five starting players on a basketball team: 71, 75, 75, 76, and 80. Move the cursor to the space under `L1`, type in 71, and hit `ENTER` (or the down arrow). There's the first player. Now enter the data for the rest of the team.

**TO CHANGE A DATUM:** Suppose the 76″ player grew since last season; his height should be listed as 78″. Use the arrow keys to move the cursor onto the 76, then change the value and `ENTER` the correction.

**TO ADD MORE DATA:** We want to include the sixth man, 73″ tall. It would be easy to simply add this new datum to the end of the list. However, sometimes the order of the data matters, so let's place this datum in numerical order. Move the cursor to the desired position (atop the first 75). Hit `2ND INS` (for "insert"), then `ENTER` the 73 in the new space.

**TO DELETE A DATUM:** The 78″ player just quit the team. Move the cursor there. Hit `DEL`. Bye.

**TO CLEAR THE DATALIST:** Finished playing basketball? Move the cursor atop the `L1`. Hit `CLEAR`, then `ENTER` (or down arrow). You should now have a blank datalist, ready for you to enter your next set of values.

**LOST A DATALIST?** Oops! Is `L1` now missing entirely? Did you delete `L1` by mistake, instead of just *clearing* it? Easy problem to fix: buy a new calculator. No? OK, then simply go to the `STAT EDIT` menu, and run `SetUpEditor` to return lists `L1` through `L6` to the `STAT EDIT` screen.

## WHAT HAVE WE LEARNED?

We've learned that data are information in a context.

- The W's help nail down the context *Who*, *What*, *When*, *Why*, *Where*, and *hoW*.
- We must know at least the *Who*, *What*, and *hoW* to be able to say anything useful based on the data. The *Who* are the cases. The *What* are the variables—the measurements made on each case. The *hoW*—how the data were collected—helps us evaluate the trustworthiness of the data.

We usually treat variables in one of two basic ways: as *categorical* or *quantitative*.

- Categorical variables identify a category for each case. Usually, we think about the counts of cases that fall into each category. (An exception is an identifier variable that just names each case.)
- Quantitative variables record measurements or amounts of something; they must have *units*.
- Sometimes we treat a variable as categorical or quantitative depending on what we want to learn from it.

### TERMS

**Data** Systematically recorded information, whether numbers or labels, together with its context. (p. 1)

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

**Context** The context ideally tells *Who* was measured, *What* was measured, *How* the data were collected, *Where* the data were collected, and *When* and *Why* the study was performed. (p. 3)

**Case** A case is an individual about whom or which we have data. (*Who*). (p. 4)

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

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

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

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

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

**Variable** A variable holds information about the same characteristic for many cases. (*What*). (p. 4)

**Sample** The cases we actually examine in seeking to understand the larger population. (p. 4)

**Population** All the cases we wish we knew about. (p. 4)

**Identifier variable** A categorical variable that assigns a unique value for each case, used to name or identify it. (p. 5)

**Categorical variable** A variable that names categories (whether with words or numerals) is called *categorical*. (p. 5)

**Quantitative variable** A variable in which the numbers act as numerical values is called *quantitative*. Quantitative variables always have units. (p. 6)

**Units** A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams. (p. 6)