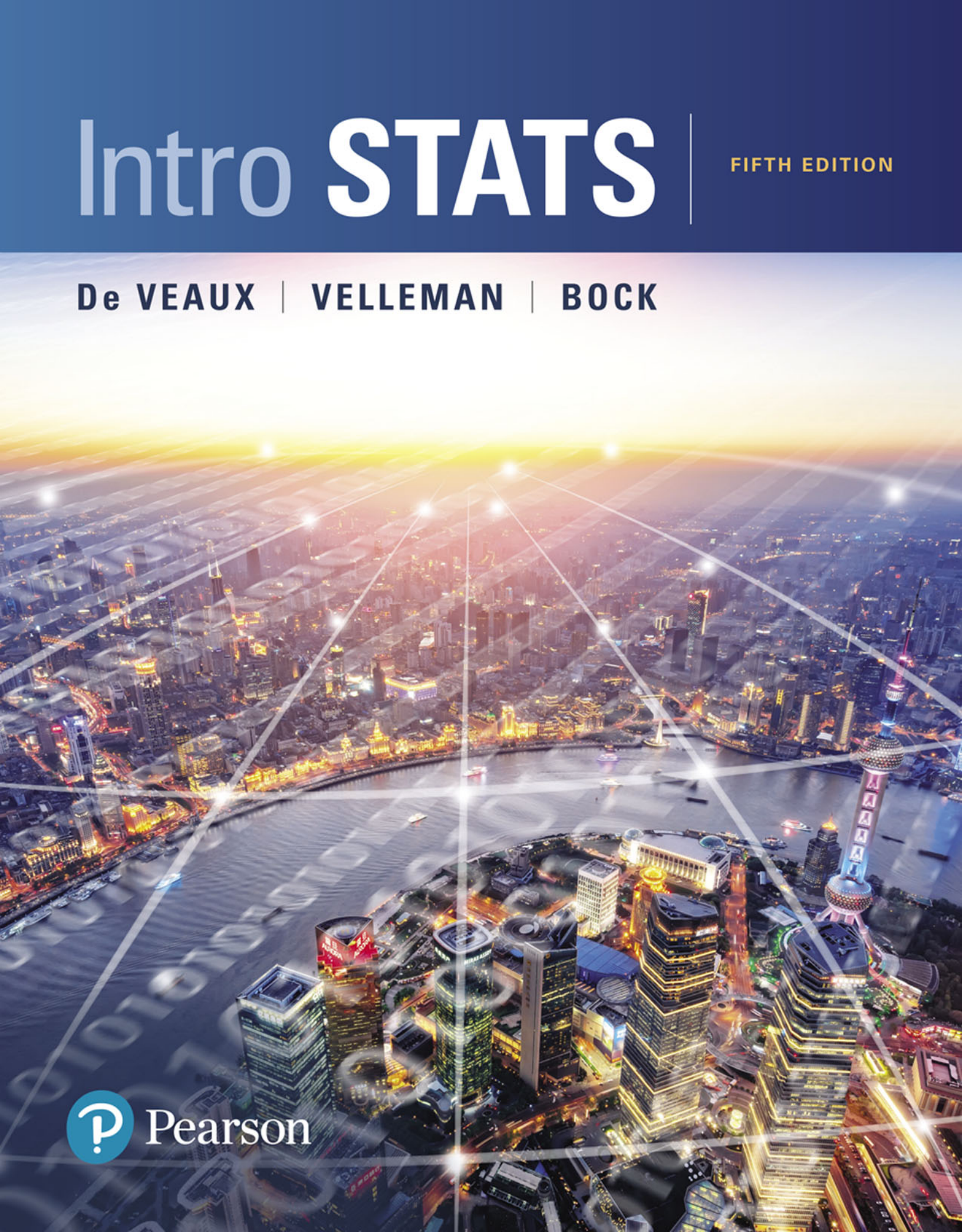


Intro **STATS**

FIFTH EDITION

De **VEAUX** | **VELLEMAN** | **BOCK**



Get the Most Out of MyStatLab[®]

MyStatLab, 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 MyStatLab into your course for the best results.

Learning in Any Environment

Because classroom formats and student needs continually change and evolve, MyStatLab has built-in flexibility to accommodate various course designs and formats.

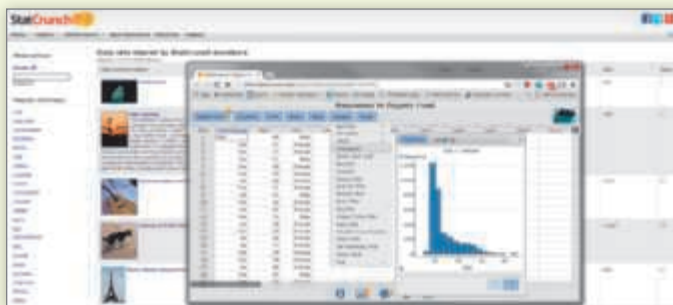
With a new, streamlined, mobile-friendly design, students and instructors can access courses from most mobile devices to work on exercises and review completed assignments.



StatCrunch

Integrated directly into MyStatLab, StatCrunch is powerful, web-based statistical software that allows students to quickly and easily analyze data sets from their text and exercises.

In addition, MyStatLab includes access to www.StatCrunch.com, the full web-based program where users can access tens of thousands of shared data sets, create and conduct online surveys, interact with applets and simulations, and perform complex analyses using the powerful statistical software.

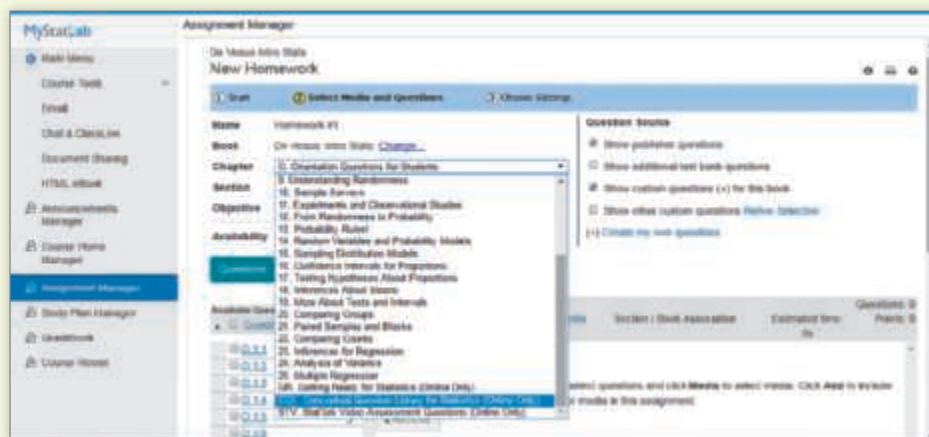


Visit www.mystatlab.com and click Training & Support to make sure you're getting the most out of MyStatLab.



Conceptual Understanding

Students need to be equipped with both the methods and conceptual understanding of statistics. MyStatLab offers a full question library of over 1,000 conceptual-based questions to help tighten the comprehension of statistical concepts.



Real-World Statistics

MyStatLab video resources help foster conceptual understanding. StatTalk Videos, hosted by fun-loving statistician, Andrew Vickers, demonstrate important statistical concepts through interesting stories and real-life events. This series of 24 videos includes assignable questions built in MyStatLab and an instructor's guide.



Visit www.mystatlab.com and click Training & Support to make sure you're getting the most out of MyStatLab.

FIFTH EDITION

Intro Stats

Richard D. De Veaux

Williams College

Paul F. Velleman

Cornell University

David E. Bock

Cornell University



Director, Portfolio Management: *Deirdre Lynch*
Courseware Portfolio Manager: *Patrick Barbera*
Content Producer: *Sherry Berg*
Managing Producer: *Karen Wernholm*
Senior Producer: *Stephanie Green*
Manager, Courseware QA: *Mary Durnwald*
Manager, Content Development: *Bob Carroll*
Product Marketing Manager: *Emily Ockay*
Field Marketing Manager: *Andrew Noble*
Marketing Assistants: *Jennifer Myers, Erin Rush*
Senior Author Support/Technology Specialist: *Joe Vetere*
Manager, Rights and Permissions: *Gina Cheselka*
Manufacturing Buyer: *Carol Melville, LSC Communications*
Art Director: *Barbara Atkinson*
Production Coordination, Composition, and Illustrations: *Cenveo® Publisher Services*
Cover Image: *Liping/Shutterstock*

Copyright © 2018, 2014, 2012 by Pearson Education, Inc. All Rights Reserved. Printed in the United States of America. This publication is protected by copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise. For information regarding permissions, request forms and the appropriate contacts within the Pearson Education Global Rights & Permissions department, please visit www.pearsoned.com/permissions/.

Attributions of third-party content appear on pages A-47–A-50, which constitutes an extension of this copyright page.

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

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

Library of Congress Cataloging-in-Publication Data

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

Title: Intro stats.

Description: Fifth edition / Richard D. De Veaux, Williams College, Paul F.

Velleman, Cornell University, David E. Bock, Cornell University

| Boston: Pearson, [2018] | Includes index.

Identifiers: LCCN [2016016733](https://lccn.loc.gov/2016016733) | ISBN 9780134210223 ((hardcover)) |

ISBN 0134210220 ((hardcover))

Subjects: LCSH: Statistics—Textbooks.

Classification: LCC QA276.12.D4 2018 | DDC 519.5—dc23

LC record available at <https://lccn.loc.gov/2016016733>

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

—Dick

*To my sons, David and Zev, from whom I've learned so much,
and to my wife, Sue, for taking a chance on me*

—Paul

*To Greg and Becca, great fun as kids and great friends as adults,
and especially to my wife and best friend, Joanna, for her
understanding, encouragement, and love*

—Dave

MEET THE AUTHORS



Richard D. De Veaux is an internationally known educator and consultant. He has taught at the Wharton School and the Princeton University School of Engineering, where he won a “Lifetime Award for Dedication and Excellence in Teaching.” He is the C. Carlisle and M. Tippet Professor of Statistics at Williams College, where he has taught since 1994. Dick has won both the Wilcoxon and Shewell awards from the American Society for Quality. He is a fellow of the American Statistical Association (ASA) and an elected member of the International Statistical Institute (ISI). In 2008, he was named Statistician of the Year by the Boston Chapter of the ASA. Dick is also well known in industry, where for more than 30 years he has consulted for such Fortune 500 companies as American Express, Hewlett-Packard, Alcoa, DuPont, Pillsbury, General Electric, and Chemical Bank. Because he consulted with Mickey Hart on his book *Planet Drum*, he has also sometimes been called the “Official Statistician for the Grateful Dead.” His real-world experiences and anecdotes illustrate many of this book’s chapters.

Dick holds degrees from Princeton University in Civil Engineering (B.S.E.) and Mathematics (A.B.) and from Stanford University in Dance Education (M.A.) and Statistics (Ph.D.), where he studied dance with Inga Weiss and Statistics with Persi Diaconis. His research focuses on the analysis of large data sets and data mining in science and industry.

In his spare time, he is an avid cyclist and swimmer. He also is the founder of the “Diminished Faculty,” an a cappella Doo-Wop quartet at Williams College and sings bass in the college concert choir and with the Chœur Vittoria of Paris. Dick is the father of four children.



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

Paul has taught Statistics at Cornell University since 1975, where he was awarded the MacIntyre Award 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. His research often deals with statistical graphics and data analysis methods. Paul co-authored (with David Hoaglin) *ABCs of Exploratory Data Analysis*. Paul is a Fellow of the American Statistical Association and of the American Association for the Advancement of Science. Paul is the father of two boys.



David E. Bock taught mathematics at Ithaca High School for 35 years. He has taught Statistics at Ithaca High School, Tompkins-Cortland Community College, Ithaca College, and Cornell University. Dave has won numerous teaching awards, including the MAA’s Edyth May Sliffe Award for Distinguished High School Mathematics Teaching (twice), Cornell University’s Outstanding Educator Award (three times), and has been a finalist for New York State Teacher of the Year.

Dave holds degrees from the University at Albany in Mathematics (B.A.) and Statistics/Education (M.S.). Dave has been a reader and table leader for the AP Statistics exam, serves as a Statistics consultant to the College Board, and leads workshops and institutes for AP Statistics teachers. He has served as K–12 Education and Outreach Coordinator and a senior lecturer for the Mathematics Department at Cornell University. His understanding of how students learn informs much of this book’s approach.

Dave and his wife relax by biking or hiking, spending much of their free time in Canada, the Rockies, or the Blue Ridge Mountains. They have a son, a daughter, and four grandchildren.

Preface ix

Index of Applications xxii

PART I Exploring and Understanding Data

1 Stats Starts Here 1

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

2 Displaying and Describing Data 17

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

3 Relationships Between Categorical Variables—Contingency Tables 64

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

4 Understanding and Comparing Distributions 95

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

5 The Standard Deviation as a Ruler and the Normal Model 122

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

Review of Part I: Exploring and Understanding Data 154

PART II Exploring Relationships Between Variables

6 Scatterplots, Association, and Correlation 163

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

7 Linear Regression 196

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

*Indicates optional sections.

8 Regression Wisdom 233

8.1 Examining Residuals ♦ **8.2** Extrapolation: Reaching Beyond the Data ♦ **8.3** Outliers, Leverage, and Influence ♦ **8.4** Lurking Variables and Causation ♦ **8.5** Working with Summary Values ♦ ***8.6** Straightening Scatterplots—The Three Goals ♦ ***8.7** Finding a Good Re-Expression

9 Multiple Regression 276

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

Review of Part II: Exploring Relationships Between Variables 307

PART III Gathering Data**10 Sample Surveys 319**

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

11 Experiments and Observational Studies 343

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

Review of Part III: Gathering Data 367

PART IV From the Data at Hand to the World at Large**12 From Randomness to Probability 373**

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

13 Sampling Distribution Models and Confidence Intervals for Proportions 410

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

14 Confidence Intervals for Means 441

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

15 Testing Hypotheses 472

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

16 More About Tests and Intervals 507

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

Review of Part IV: From the Data at Hand to the World at Large 532**PART V Inference for Relationships****17 Comparing Groups 541**

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

18 Paired Samples and Blocks 585

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

19 Comparing Counts 610

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

20 Inferences for Regression 642

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

Review of Part V: Inference for Relationships 685**Parts I–V Cumulative Review Exercises 698****Appendixes**

A Answers **A-1** ♦ **B** Credits **A-47** ♦ **C** Indexes **A-51**
♦ **D** Tables and Selected Formulas **A-63**

This page intentionally left blank

Intro Stats, 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 include models with more than two variables. We've taken our inspiration both from our experience in the classroom and from the 2016 revision of the Guidelines for Assessment and Instruction in Statistics Education (GAISE) report adopted by the American Statistical Association. As a result, we increased the text's innovative uses of technology to encourage more statistical thinking, while maintaining its traditional core concepts and coverage. You'll notice that, 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 data sets. Throughout, you'll find a focus on statistical thinking rather than calculation. The question that motivates each of our hundreds of examples is not “How do you calculate the answer?” but “How do you think about the answer?”

For this edition of *Intro Stats* we've taken this principle still further. We have harnessed technology to 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 *Intro Stats* 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 is worth the effort. We've added new material about interpreting what regression models say. The effectiveness of multiple regression is immediately obvious and makes the reach and power of statistics clear. The use of real data underscores the universal applicability of these methods.

When we return to regression in Chapter 20 to discuss inference, we can deal with both simple and multiple regression models together. There is nothing different to discuss.

¹And footnotes

(For this reason we set aside the F -test and adjusted R^2 . Students can add those later if they need them.) This course is an *introduction* to statistics. It isn't necessary to learn *all* the details of the methods and models. But it is important to come away with a sense of the power and usefulness of statistics to solve real problems.

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

Bootstrap

The introduction to the new GAISE guidelines explicitly mentions the bootstrap method. The bootstrap is not as widely available or as widely understood as multiple regression. But it 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 re-sampling from the sample at hand.

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

Real Data

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

Let us be clear. *Intro Stats* comes with an archive of nearly 300 datasets used in more than 600 applications throughout the book. The datasets are available online at the student resource site and in MyStatlab. 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 discussion of random variables and probability distributions is shorter than in previous editions—again, a GAISE recommendation. Those are interesting topics, but they are not needed in this course. We leave them for a later course for those students who want to go further.

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

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

GAISE 2016

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

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

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

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

Chapter Order

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

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

We’ve found that one of the challenges students face is how to know what technique to use when. In the real world, questions don’t come at the ends of the chapters. So, as always, we’ve provided summaries at the end of each part along with a series of exercises

designed to stretch student understanding. These Part Reviews are a mix of questions from all the chapters in that part. Finally, we've added an extra set of "book-level" review problems at the end of the book. These ask students to integrate what they've learned from the entire course. The questions range from simple questions about what method to use in various situations to 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 20.)
- ◆ **The need to read.** Statistics is a consistent story about how to understand the world when we have data. The story can't be told piecemeal. This is a book that needs to be read, so we've tried to make the reading experience enjoyable. Students who start with the exercises and then search back for a worked example that looks the same but with different numbers will find that our presentation doesn't support that approach.

Mathematics

Mathematics can make discussions of statistics concepts, probability, and inference clear and concise. We don't shy away from using math where it can clarify without intimidating. But we know that some students are discouraged by equations, so we always provide a verbal description and a numerical example as well.

Nor do we slide in the opposite direction and concentrate on calculation. Although statistics calculations are generally straightforward, they are also usually tedious. And, more to the point, today, virtually all statistics are calculated with technology. We have selected the equations that focus on illuminating concepts and methods rather than for hand calculation. We sometimes give an alternative formula, better suited for hand calculation, for those who find that following the calculation process is a better way to learn about the result.

Technology and Data

We assume that computers and appropriate software are available—at least for demonstration purposes. We hope that students have access to computers and statistics software for their analyses.

We discuss generic computer output at the end of most chapters, but we don't adopt any particular statistics software. The **Tech Support** sections at the ends of chapters offer guidance for seven common software platforms: Data Desk, Excel, JMP, Minitab, SPSS,

StatCrunch, and R. We also offer some advice for TI-83/84 Plus graphing calculators, although we hope that those who use them will also have some access to computers and statistics software.

We don't limit ourselves to small, artificial data sets, but base most examples and exercises on real data with a moderate number of cases. Machine-readable versions of the data are available at the book's website, pearsonhighered.com/dvb.

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.

- ◆ **Connections** specifically ties the new topics to those learned in previous chapters.
- ◆ **What Can Go Wrong?** sections highlight the most common errors that people make and the misconceptions they have about statistics. One of our goals is to arm students with the tools to detect statistical errors and to offer practice in debunking misuses of statistics, whether intentional or not.
- ◆ Next, the **Chapter Review** summarizes the story told by the chapter and provides a bullet list of the major concepts and principles covered.
- ◆ A **Review of Terms** is a glossary of all of the special terms introduced in the chapter. In the text, these are printed in **bold** and underlined. The Review provides page references, so students can easily turn back to a full discussion of the term if the brief definition isn't sufficient.

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

Learning by Example



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

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

Additional Resources Online

Most of the supporting materials can be found online:

At the book's website at pearsonhighered.com/dvb

Within the MyStatlab course at www.mystatlab.com

Datasets are also available at dasl.datadesk.com.

Data desk 8 is a statistics program with a graphical interface that is easy to learn and use. A student version is available at datadesk.com. Click on the **Teachers & Students** tab at the top of the page.

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.

MyStatLab™ Online Course (access code required)

MyStatLab from Pearson is the world's leading online resource for teaching and learning statistics; integrating interactive homework, assessment, and media in a flexible, easy-to-use format. It is a course management system that delivers proven results in helping individual students succeed.

- ◆ MyStatLab can be successfully implemented in any environment—lab-based, hybrid, fully online, traditional—and demonstrates the quantifiable difference that integrated usage has on student retention, subsequent success, and overall achievement.
- ◆ MyStatLab's comprehensive online gradebook automatically tracks students' results on tests, quizzes, homework, and in the study plan. Instructors can use the gradebook to provide positive feedback or intervene if students have trouble. Gradebook data can be easily exported to a variety of spreadsheet programs, such as Microsoft Excel.

MyStatLab provides engaging experiences that personalize, stimulate, and measure learning for each student. In addition to the resources below, each course includes a full interactive online version of the accompanying textbook.

- ◆ **Personalized Learning:** MyStatLab's personalized homework, and adaptive and companion study plan features allow your students to work more efficiently, spending time where they really need to.
- ◆ **Tutorial Exercises with Multimedia Learning Aids:** The homework and practice exercises in MyStatLab align with the exercises in the textbook, and they regenerate algorithmically to give students unlimited opportunity for practice and mastery. Exercises offer immediate helpful feedback, guided solutions, sample problems, animations, videos, and eText clips for extra help at point-of-use.
- ◆ **Learning Catalytics™:** MyStatLab now provides Learning Catalytics—an interactive student response tool that uses students' smartphones, tablets, or laptops to engage them in more sophisticated tasks and thinking.
- ◆ **Getting Ready for Statistics:** A library of questions now appears within each MyStatLab course to offer the developmental math topics students need for the course. These can be assigned as a prerequisite to other assignments.
- ◆ **Conceptual Question Library:** A library of 1,000 Conceptual Questions available in the assignment manager requires students to apply their statistical understanding.
- ◆ **StatTalk Videos:** Fun-loving statistician Andrew Vickers takes to the streets of Brooklyn, NY, to demonstrate important statistical concepts through interesting stories and real-life events. This series of 24 fun and engaging videos will help students actually understand statistical concepts. Available with an instructor's user guide and assessment questions.
- ◆ **StatCrunch™:** MyStatLab integrates the web-based statistical software, StatCrunch, within the online assessment platform so that students can easily analyze data sets from exercises and the text. In addition, MyStatLab includes access to www.statcrunch.com, a vibrant online community where users can access tens of thousands of shared data sets, create and conduct online surveys, perform complex analyses using the powerful statistical software, and generate compelling reports.
- ◆ **Statistical Software Support and Integration:** We make it easy to copy our data sets, both from the ebook and the MyStatLab questions, into software such as StatCrunch, Minitab, Excel, and more. Students have access to a variety of support tools—Tutorial Videos, Technology Study Cards, and Technology Manuals for select titles—to learn how to effectively use statistical software.
- ◆ **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.

MathXL[®] for Statistics Online Course (access code required)

Part of the world's leading collection of online homework, tutorial, and assessment products, Pearson MathXL delivers assessment and tutorials 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 website at www.pearson.com/mathxl, or contact your Pearson representative.

StatCrunch[™]

StatCrunch is powerful web-based statistical software that allows users to perform complex analyses, share data sets, and generate compelling reports of their data. The vibrant online community offers tens of thousands shared data sets for students to analyze.

- ◆ **Collect.** Users can upload their own data to StatCrunch or search a large library of publicly shared data sets, spanning almost any topic of interest. Also, an online survey tool allows users to quickly collect data via web-based surveys.
- ◆ **Crunch.** A full range of numerical and graphical methods allow users to analyze and gain insights from any data set. Interactive graphics help users understand statistical concepts and are available for export to enrich reports with visual representations of data.
- ◆ **Communicate.** Reporting options help users create a wide variety of visually appealing representations of their data.

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

Additional Resources

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

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



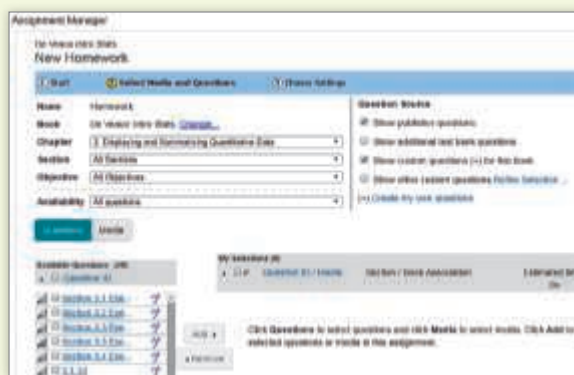
Resources for Success

MyStatLab® Online Course for *Intro Stats, 5e*
by Richard D. De Veaux, Paul F. Velleman,
and David E. Bock (access code required)

MyStatLab 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 MyStatLab course, making learning the material as seamless as possible.

Expanded objective-based exercise coverage

MyStatLab exercises are newly mapped to improve student learning outcomes. Homework reinforces and supports students’ understanding of key statistics topics.



College Students

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

Enhanced video program to meet Introductory Statistics needs

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

Examples and exercises throughout the textbook and MyStatLab use current, real-world data to help students understand how statistics applies to everyday life.

Movie_budgets.txt

Rank	Movie	Run Time (minutes)	Budget (M\$)	Grosses (M\$)	Genre	Year
1	White Noise	101	30	58094360	Horror	0
2	Coach Carter	130	45	67364817	Drama	1
3	Warrior	100	65	34409722	Action	0
4	Raising Helen	110	30	48773222	Comedy	0
5	Support Your Local Lady	105	30	38540468	Action	0
6	Are We There Yet?	94	30	82674398	Comedy	0
7	Along with the Gods	96	30	8176259	Horror	0
8	Madagascar	100	29	18100488	Drama	1
9	Empire of the Sun	99	30	46753282	Horror	0
10	The Wedding Date	90	40	21708899	Comedy	0
11	Mr. Bean	113	30	177982098	Comedy	0
12	Paul's HulaHoop Movie	68	30	18281626	Adventure	0
13	Revenge of Wer-Oh!	106	10	22647042	Comedy	0
14	Conanoban	100	75	79576178	Action	0
15	Salt of the Earth	96	100	17018422	Comedy	0
16	Carters	100	30	18294960	Horror	0
17	Wall of the Heavens	100	30	38697706	Comedy	0
18	Be Cool	150	30	22666401	Comedy	0
19	The Jockey	103	30.0	6301131	Thriller	0
20	The Pacifier	97	56.3	130000906	Comedy	0
21	Hatigga	108	75	28628442	Action	0
22	Robots	90	60	126000108	Adventure	0
23	The Upside of Anger	117	32	18761993	Drama	1
24	Joe Verses the Volcano	99	25	18181214	Comedy	0
25	The Ring Two	111	58	75643727	Horror	0
26	Hot Chick 2: Armed & Dangerous	113	60	48476626	Comedy	0



Resources for Success

Student Resources

Intro Stats, 5th edition is part of De Veaux, Velleman, and Bock's Statistics series (ISBN-13: 978-0-13-421022-3; ISBN-10: 0-13-421022-0)

Student's Solutions Manual by William Craine, provides detailed, worked-out solutions to odd-numbered exercises. This manual is available within MyStatLab. (ISBN-13: 978-0-13-426535-3; ISBN-10: 0-13-426535-1)

Instructor Resources

Instructor's Edition contains answers to all exercises, plus recommended assignments and teaching suggestions. (ISBN-13: 978-0-13-421036-0; ISBN-10: 0-13-421036-0)

Instructor's Solutions Manual (Download Only), by William Craine, contains solutions to all the exercises. These files are available to qualified instructors through Pearson Education's online catalog at www.pearsonhighered.com/irc or within MyStatLab.

Online Test Bank and Resource Guide (Download Only), by 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. These files are available to qualified instructors through Pearson Education's online catalog at www.pearsonhighered.com/irc or within MyStatLab.

TestGen® Computerized Test Bank

(www.pearsoned.com/testgen) enables instructors to build, edit, print, and administer tests using a computerized bank of questions developed to cover all the objectives of the text. TestGen is algorithmically based, allowing instructors to create multiple but equivalent versions of the same question or test with the click of a button. Instructors can also modify test bank questions or add new questions. The software and test bank are available for download from Pearson Education's online catalog at www.pearsonhighered.com.

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

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

ACKNOWLEDGMENTS

Many people have contributed to this book throughout all of its editions. This edition never would have seen the light of day without the assistance of the incredible team at Pearson. Director, 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. Sherry Berg, 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. Justin Billing, Portfolio Management Assistant, Jennifer Myers, Marketing Assistant, and Erin Rush, Field Marketing Assistant, were essential in managing all of the behind-the-scenes work. Senior Producer Stephanie Green put together a top-notch media package for this book. Senior Project Manager Chere Bemelmans of Cenveo Publisher Services 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, to 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 data sets.

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 <i>Cornell College</i>	Sheldon Lee <i>Viterbo University</i>	Dirk Tempelaar <i>Maastricht University</i>
Susan Chimiak <i>University of Maryland</i>	Pam Omer <i>Western New England University</i>	Carol Weideman <i>St. Petersburg College</i>
Lynda Hollingsworth <i>Northwest Missouri State University</i>	Sarah Quesen <i>West Virginia University</i>	Ming Wang <i>University of Kansas</i>
Jeff Kollath <i>Oregon State University</i>	Karin Reinhold <i>SUNY Albany</i>	Lisa Wellinghoff <i>Wright State</i>
Cindy Leary <i>University of Montana</i>	Laura Shick <i>Clemson University</i>	Cathy Zucco-Teveloff <i>Rider University</i>

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

Mary Kay Abbey <i>Montgomery College</i>	Sanjib Basu <i>Northern Illinois University</i>	Jerry Chen <i>Suffolk County Community College</i>
Froozan Pourboghna Afiat <i>Community College of Southern Nevada</i>	Carl D. Bodenschatz <i>University of Pittsburgh</i>	Rick Denman <i>Southwestern University</i>
Mehdi Afiat <i>Community College of Southern Nevada</i>	Steven Bogart <i>Shoreline Community College</i>	Jeffrey Eldridge <i>Edmonds Community College</i>
Nazanin Azarnia <i>Santa Fe Community College</i>	Ann Cannon <i>Cornell College</i>	Karen Estes <i>St. Petersburg Junior College</i>
	Robert L. Carson <i>Hagerstown Community College</i>	

- Richard Friary
Kim (Robinson) Gilbert
Clayton College & State University
- Ken Grace
Anoka-Ramsey Community College
- Jonathan Graham
University of Montana
- Nancy Heckman
University of British Columbia
- James Helreich
Marist College
- Susan Herring
Sonoma State University
- Mary R. Hudachek-Buswell
Clayton State University
- Patricia Humphrey
Georgia Southern University
- Becky Hurley
Rockingham Community College
- Debra Ingram
Arkansas State University
- Joseph Kupresanin
Cecil College
- Kelly Jackson
Camden County College
- Martin Jones
College of Charleston
- Rebecka Jornsten
Rutgers University
- Michael Kinter
Cuesta College
- Kathleen Kone
Community College of Allegheny County
- Michael Lichter
State University of New York–Buffalo
- Susan Loch
University of Minnesota
- Pamela Lockwood
Western Texas A & M University
- Wei-Yin Loh
University of Wisconsin–Madison
- Steve Marsden
Glendale College
- Catherine Matos
Clayton College & State University
- Elaine McDonald
Sonoma State University
- Jackie Miller
The Ohio State University
- Hari Mukerjee
Wichita State University
- Helen Noble
San Diego State University
- Monica Oabos
Santa Barbara City College
- Linda Obeid
Reedley College
- Charles C. Okeke
Community College of Southern Nevada
- Pamela Omer
Western New England College
- Mavis Pararai
Indiana University of Pennsylvania
- Gina Reed
Gainesville College
- Juana Sanchez
UCLA
- Gerald Schoultz
Grand Valley State University
- Jim Smart
Tallahassee Community College
- Chamont Wang
The College of New Jersey
- Edward Welsh
Westfield State College
- Heydar Zahedani
California State University, San Marcos
- Cathy Zucco-Teveloff
Rider University
- Dottie Walton
Cuyahoga Community College
- Jay Xu
Williams College

INDEX OF APPLICATIONS

BE = Boxed Example; E = Exercise; IE = In-Text Example; JC = Just Checking; RM = Random Matters; SBS = Step-by-Step examples; WCGW = What Could Go Wrong

Accounting

Troubleshooting, (E): 341

Advertising

Appliance Sales, (E): 301
Cell Phones, (JC): 421–422
Direct Mail, (E): 439
Endorsements, (E): 537
Political Ads, (E): 501, 502
Radio Advertising, (E): 529
Sexual Images in Advertising, (E): 603, 608
Super Bowl Commercials, (E): 362, 504
Television Advertising, (E): 581

Agriculture

Chicken Feed, (E): 602–603
Egg Production, (E): 153, 602–603
Insect Control, (E): 365, (JC): 623
Livestock, (E): 463, 535, 609
Potatoes, (E): 437
Seed Viability, (E): 503
Vineyards, (E): 15, 118, 308, 371

Banking

Credit Cards, (BE): 429, (E): 156, 261, 435, 436, 438, 464, 634, (JC): 493
Customer Age, (E): 634
Loans, (E): 528
Online Banking, (E): 403
Website, (E): 530

Business (General)

Assets and Sales, (E): 316–317
CEO Compensation, (BE): 130–131, (E): 151–152, 464, 467, 468, (IE): 105–106, 443–444, (WCGW): 142–143
Employee Injuries, (E): 437
Profits, (E): 316, 361
Women Executives, (E): 503
Women-Owned Firms, (E): 535, 537

Company Names

Allstate Insurance Company, (E): 468
Amazon, (BE): 4, (E): 13, (IE): 2, 3, 4, 5, 6
AT&T, (BE): 5
Bentley, (IE): 247
Cleveland Casting Plant, (E): 14
Daimler AG, (IE): 247

Facebook, (BE): 6, (E): 13, 403, 505, 506, 573, (IE): 2, 393, 423–424, (JC): 484, (SBS): 394
Ferrari World, (BE): 105
Ford Motor Company, (E): 14
Fukushima Daiichi Nuclear Power Plant, (BE): 24–25
GfK Roper, (E): 339, 576
GlaxoSmithKline (GSK), (BE): 518, 522
Google, (BE): 6
Guinness Company, (IE): 445–446
Lowe's, (IE): 242
Mayo Clinic, (IE): 410
Nabisco Company, (E): 469
OkCupid, (IE): 64, 65–66, 78
Preusser Group, (IE): 507
Rolls-Royce, (IE): 247
SmartWool, (BE): 474, (IE): 475–476
Summit Projects, (BE): 474
Twitter, (IE): 393, (SBS): 394
White Star Line, (IE): 34

Consumers

Assembly Time, (E): 162
Consumer Price Index, (BE): 29–30
Credit Card Expenditures, (BE): 30, 35, 38
Pet Ownership, (E): 403
Purchases, (IE): 551, (SBS): 556–558
Shoes, (E): 147, 464
Tipping, (E): 313, 361, 369, 462
Wardrobe, (E): 402

Demographics

Age of Couples, (IE): 592, 595
Deaths, (E): 116
Egyptian Body Structure, (E): 579
Foreign-Born Citizens, (E): 572
Marital Status/Age, (BE): 103, 238–239, (E): 52, 116, 263, 267, 268, 269, 435, 501, 535, 676, 677
Population, (E): 60, 62, 115, (IE): 176–177, (JC): 250, 450–451, 453
Poverty, (E): 87
Race/Ethnicity, (E): 151, 501
U.S. Census, (E): 501

Distribution and Operations Management

Delivery Services and Times, (E): 93
Shipments, (E): 147

E-Commerce

Online Shopping, (E): 13, 188, 225
Profits, (E): 361

Economics

Cost of Living, (BE): 165–166, 172, 175, (E): 114, 121, 228
Crowdedness, (E): 271, 272
GDP, (E): 272, 273
Gross Domestic Product, (E): 190–191
Human Development Index (HDI), (E): 264
Income, (JC): 37
Income and Housing Costs, (E): 191
Inflation, (E): 270
Labor Force, (E): 86
Labor Force Participation Rate, (E): 603
Market Segments, (E): 261
Occupy Wall Street Movement, (E): 538
Prices, (JC): 250
United Nations Development Programme (UNDP), (E): 264

Education

Academic Performance, (IE): 343–344
Admissions/Placement, (BE): 80–81, (E): 93, 147, 404
Attainment by Age Group, (E): 640
Cheating, (E): 536
College Attendance Rates, (E): 501
College Retention rates, (E): 439, 440
College Value, (E): 85–86, 93
Core Plus Mathematics Project (CPMP), (E): 577
Cornell University, (IE): 355
Cost of Higher Education, (E): 607, 684
Cramming, (E): 156, 310
Dartmouth College, (E): 464
Earnings of College Graduates, (E): 672, 673, 674
Educational Testing Service (ETS), (JC): 127
Employment of College Students, (IE): 492
Grade Levels, (E): 13
Grades/Scores/GPA, (BE): 392, 658, (E): 58, 61, 62, 117, 118, 119, 146, 147, 148, 149, 150, 152, 225, 226, 262, 266, 301, 303, 304, 305, 308, 310, 314, 340, 361, 364, 365, 436, 462, 468, 500, 501, 504, 534, 537, 639, 640, 681–682, (IE): 135, 447, (JC): 125, 127, 153, 172–173, 195, (SBS): 128, 131–132, 136, 137

Graduation Rates, (E): 117, 408, 437, 440
 Group Projects, (E): 405
 Height and Reading, (E): 191
 High-School Dropout Rate, (E): 504
 High-School Graduation Rate, (E): 574
 Homecoming, (E): 370
 Kindergarten, (E): 153
 Living arrangements, (E): 635, (JC): 629
 LSAT, (E): 462
 Magnet Schools, (E): 56, 87, 90
 Major Field, (E): 160
 Meal Plans, (E): 464
 Mortality and Education, (E): 683
 National Center for Education Statistics, (E): 503
 Parental education, (E): 503
 Post-Graduation Activities, (E): 89, 90, (IE): 618–620, 622, (SBS): 620–621
 Prerequisites, (E): 407
 Professors, (E): 150
 Public Opinion, (E): 155
 Reading, (E): 266, 365, 368, 471, 579, (IE): 527
 School Absenteeism, (E): 503
 School Mail, (E): 160
 School Records, (E): 14
 School Uniforms, (E): 437
 Software, (E): 368, 529
 Spring Break, (E): 435
 Student Evaluations, (IE): 355–356
 Student Goals, (IE): 389–390
 Studying, (E): 160
 Summer School, (E): 580, 606
 Teachers, (E): 439
 Texas A&M University, (E): 338
 University of California at Berkeley, (BE): 80–81, (E): 93
 University of Texas, (E): 315
 Williams College, (E): 315, (IE): 95, 104

Energy

Batteries, (E): 407, 539
 Energy Information Administration (EIA), (IE): 238
 Fuel Economy, (E): 15, 116–117, 119, 147, 191, 192, 221, 231, 271, 272, 275, 311–312, 341, 365, 371, 602, 606–607, 678, 679, 682–683, (IE): 141–142, 246–248, 281, (JC): 37, (SBS): 38–39
 Fukushima Daiichi Nuclear Power Plant, (BE): 24–25
 Gas Prices, (E): 60, 116
 Oil Prices, (IE): 237–238
 Wind Power, (E): 470–471

Environment

Acid Rain, (E): 61, 155–156, 504
 Camping, (E): 115
 City Climate, (E): 680

Cloud Seeding, (E): 119, 120
 Earthquakes and Tsunamis, (BE): 24–25
 Emissions Testing, (E): 440, 528–529
 Environmental Protection Agency, (E): 15, 147
 Floods, (E): 58
 Global Warming, (BE): 428, (E): 13, 14, 57, 229, 368, 678, 679
 Hard Water, (E): 580, 583
 Hurricanes, (BE): 199, 206, 214, (E): 60, 190, 265, 636, 674, (IE): 104, 163–165, (JC): 287, 306
 National Hurricane Center (NHC), (IE): 163, 167
 National Oceanic and Atmospheric Administration (NOAA), (IE): 163
 National Weather Service, (IE): 104
 Northeast Regional Climate Center, (E): 469
 Old Faithful, (E): 158, 312
 Ozone, (E): 117, 680
 Pollution, (E): 503
 Rainfall, (E): 469, 603
 Sea Ice, (E): 674–675
 Seasons, (E): 157–158
 Snow, (E): 464
 Soil, (E): 339
 Streams, (E): 15, 156, 190, 309, 310, 538, 579, 677–678
 Temperature, (E): 148, 190, 266, 313, 605, 680
 Tornadoes, (E): 58
 Typhoons, (IE): 163
 Water, (E): 157, 230
 Weather, (BE): 199, 206, 214, (E): 13, 60, 91, 146, 162, 188, 190, 265, 403–404, 464, 469, 636, 674, (IE): 104, 163–165, (JC): 287
 Wildfires, (E): 226
 Wind Speed, (E): 149, 271, 300, 604, 605, (IE): 95–99, 103, 104, (JC): 306

Famous People

Archimedes, (IE): 241
 Armstrong, Lance, (JC): 7
 Bacon, Francis, (IE): 241, 648
 Bayes, Thomas, (BE): 399, (IE): 398
 Bernoulli, Jacob, (IE): 375
 Berra, Yogi, (IE): 164, 380
 Bohr, Neils, (IE): 236
 Box, George, (IE): 131, 197
 Boyle, Robert, (E): 273
 Brahe, Tycho, (IE): 8
 Buchanan, Pat, (IE): 240–241
 Bush, George W., (IE): 240–241, 319, 332
 Carroll, Lewis, (IE): 1, 395
 Ceci, Stephen, (IE): 355–356
 Clinton, Bill, (E): 537
 Clinton, Hilary, (E): 534
 Collier, Wayne, (E): 371
 Cornet, Henri, (JC): 7

de Moivre, Abraham, (IE): 129, 412, 414
 Descartes, René, (BE): 166
 Efron, Bradley, (BE): 454
 Einstein, Albert, (E): 465
 Evans, Cadel, (JC): 7
 Fechner, Gustav, (IE): 345
 Fisher, Ronald Aylmer, (BE): 177, 512, 518, (IE): 446n, 449, 476
 Fleet, Frederick, (IE): 17
 Froome, Christopher, (JC): 7
 Gallup, George, (IE): 321
 Galton, Francis, (BE): 204, (IE): 205
 Garin, Maurice, (JC): 7
 Gauss, Carl Friedrich, (BE): 198, (IE): 129, 446n
 Gill, Colin, (E): 370
 Gore, Al, (IE): 240–241
 Gosset, W. S., (IE): 445–446, 447
 Gretzky, Wayne, (E): 60
 Halifax, Lord, (IE): 344
 Harvey, William, (IE): 240
 Homer, (E): 635
 Howe, Gordie, (E): 60
 Hume, David, (IE): 510
 Jastrow, J., (IE): 346
 Johnson, Gary, (E): 534
 Johnson-Thompson, Katarina, (IE): 124, 125, 126
 Kepler, Johannes, (IE): 8, 9
 Klassen, Cindy, (SBS): 589
 Kohavi, Ronny, (BE): 4, (IE): 2
 Landers, Ann, (BE): 333
 Landon, Alf, (IE): 320, 321, 334
 Laplace, Pierre-Simon, (IE): 442
 Legendre, Adrien-Marie, (BE): 198
 Ligety, Ted, (BE): 126
 Lowell, James Russell, (IE): 480
 Maas, Jim, (IE): 484
 McGwire, Mark, (E): 159
 Meir, Jessica, (IE): 234
 Michelson, Albert Abraham, (E): 465, 466
 Miller, Bode, (BE): 126
 Moore, David, (IE): 348n
 Munchausen, Baron, (IE): 454n
 Nader, Ralph, (IE): 240, 241
 Newton, Isaac, (BE): 166, (IE): 9
 Nibali, Vincenzo, (JC): 7
 Nobel Laureates, (E): 13
 Obama, Barack, (E): 369, 527, 537
 O'Loughlin, William Francis Norman, (IE): 31
 Pierce, C. S., (IE): 346, 351n
 Poganis, Paul, (IE): 234
 Quenouille, Maurice, (BE): 454
 Raspe, Rudolf Erich, (IE): 454n
 Robbins, Rebeca, (IE): 484
 Rodriguez, Alex, (E): 60
 Roosevelt, Franklin Delano, (IE): 320, 321, 334
 Rudder, Christian, (IE): 64
 Ruth, Babe, (E): 159
 Saunderson, Nicholas, (IE): 399

Schleck, Andy, (JC): 7
 Smith, Edward John, (IE): 31
 Sophocles, (IE): 307
 Spicer, Sean, (BE): 67
 Stein, Jill, (E): 534
 Stigler, Steven, (BE): 399
 Thiam, Nafissatou, (IE): 124, 125, 126
 Trousseller, Louis, (JC): 7
 Trump, Donald J., (E): 534, (IE): 319
 Truzzi, Marchello, (IE): 492
 Tufte, Edward, (WCGW): 43n
 Tukey, John W., (BE): 26, 99, 454, (IE): 249, 421
 Van Buren, Abigail, (BE): 376
 Venn, John, (IE): 378
 Wannamaker, John, (IE): 472
 Watson, William Albert, (IE): 31
 Wayne, John, (E): 504
 Weeks, David, (E): 370
 Wiggins, Bradley, (JC): 7
 Wunderlich, Carl, (IE): 486, 487

Finance and Investments

Assets, (E): 120
 Brokerage Accounts, (E): 635
 Computer Lab Fees, (E): 466, 504–505
 Currency, (BE): 391, (E): 527, (IE): 377
 Drug Development Costs, (E): 501
 401(k) Plans, (E): 14
 Interest Rates, (E): 191, 224
 Mutual Funds, (E): 220
 Profits, (E): 162
 Saving and Investment Performance, (BE): 97
 Second Jobs, (E): 527
 Stocks, (E): 309, 534
 T Bill Rates, (E): 267, 268, 272
 Websites, (E): 500

Food/Drink

Alcohol Consumption, (E): 315, 362, 364, 370–371, 408, 536, (IE): 395
 Appetite, (E): 581
 Bananas, (E): 154
 Beer, (IE): 445–446
 Bread, (E): 155, 538
 Burgers, (E): 192
 Caffeine, (E): 118
 Candy, (E): 305, 405, 502, 635, (SBS): 386–388
 Cereal, (E): 57, 114, 162, 221, 304–305, 578, 679, (IE): 213–214, 235–236, (SBS): 138–140, 211–213
 Chicken, (E): 407
 Coffee, (E): 189
 Cookies, (E): 370, 469, 505, 530–531
 Cooking, (E): 403
 Cranberry Juice, (E): 638–639
 Diet, (E): 576
 Fast Food, (E): 576
 Fish, (BE): 448, 483–484, (E): 462, 466, 531, 639

Food Preferences, (E): 576
 Grocery Shopping, (E): 13
 Hams, (E): 148
 Hot Dogs, (E): 504, 675, 676
 Meals, (E): 315, 538, (JC): 629
 Milk, (E): 339
 Nutritional Data, (BE): 200–201, 292, (E): 227, 228, 263, 302, 466, 501, 504, 576–577, 675, 676, (IE): 196–197, 198–199, 201, 205, 206, 207–208, 210, (WCGW): 214–215
 Nuts, (E): 636
 Pet Food, (BE): 347, 352, (E): 363, 367
 Pizza, (E): 58, 59, 113, 504
 Quality Control, (BE): 347, 352, (E): 341
 Restaurants, (E): 409
 Safety, (BE): 448, 483–484, (E): 368, 439
 Salt, (E): 162
 Seafood, (E): 438
 Snack Foods, (E): 339, 469, 505, 538
 Soft Drinks, (E): 501, (IE): 527
 Soup, (BE): 552, 554–555, (E): 436, (IE): 320–322
 Thirst, (E): 581
 Tomatoes, (E): 153, 361, 363
 Unsafe Food, (E): 407
 Yogurt, (E): 469, 505, 606

Games

Cards, (E): 156, 406
 Coin Flips, (E): 502, (IE): 492
 Coin Spins, (E): 437, 530
 Coin Toss, (BE): 375–376, (E): 402
 Dice, (E): 402, 502, 635
 Gambling, (BE): 376, (E): 404, 439, 500, (IE): 659
 Keno, (BE): 376
 Lottery, (E): 406, 468, 636–637, (JC): 409
 Roulette, (E): 403
 Scrabble, (E): 534
 Spinners, (E): 404

Government, Labor, and Law

Arrests, (E): 640
 Bureau of Labor Statistics, (BE): 25
 Car Thefts, (E): 189
 Checkpoints, (E): 437
 Crime and Television Watching, (E): 189
 Death Penalty, (E): 439, (SBS): 426–427
 Juries, (E): 503, (IE): 474–475, 476
 Labor Force, (E): 86
 Labor Force Participation Rate, (E): 603
 National Security Agency (NSA), (IE): 6n
 NYPD, (E): 638
 Parks, (E): 340
 Parole, (E): 438
 Polygraphs, (E): 409
 Prisons, (E): 112, 439

Radon Testing, (E): 367
 Roadblocks, (E): 339
 Speeding, (E): 404, 537
 Taxes, (E): 404
 Traffic Stops, (BE): 624, 626–627, (E): 636
 U.S. Census Bureau, (BE): 323, (JC): 37, 450–451, 453
 U.S. Postal Service, (E): 13
 Violence against Women, (E): 636
 Zip Codes, (E): 13, 61

Human Resource Management/Personnel

Absenteeism, (E): 408
 Career Success, (E): 86, 88
 College Student Employment, (IE): 492
 Commute Times, (RM): 40–41, 133–134
 Flexible Work Schedules, (BE): 586, 588–589, 591–592
 Hiring, (E): 437, 440
 Hours Worked, (BE): 25
 Human Resource Data, (E): 533
 Job Discrimination, (E): 529
 Job Growth, (E): 62
 Job Hunting, (E): 13
 Job Satisfaction, (E): 89, 338, 339, 341, 572, 580, 606
 Labor Force, (E): 86
 On-Site Day Care, (E): 500
 Placement Scores, (IE): 527
 Rating Employees, (E): 220
 Sick Days, (E): 58
 Workplace Ratings, (E): 94

Insurance

Auto Insurance, (E): 404, 468, 604, 605
 Health Insurance, (E): 159, 407, 537–538

Manufacturing

Assembly Time, (E): 162
 Beer, (IE): 445–446
 Cars, (E): 438, 635, 639
 Catheters, (E): 533
 Chips, (E): 316
 Computers, (E): 119
 Electronics, (E): 409
 Emergency Shutoff, (E): 533
 Machine Settings, (BE): 565
 Metals, (E): 14, (IE): 472–475, 476–477, (RM): 523–524
 Pottery, (E): 189
 Safety, (E): 365
 Shoes, (E): 54, 363, (IE): 350
 Skates, (JC): 127
 Swimsuits, (E): 364

Marketing

Direct Marketing, (E): 434

Media and Entertainment

American Journal of Health Behavior, (IE): 395
Archives of General Psychiatry, (E): 575
Berkshire Eagle, (IE): 104
British Medical Journal, (E): 638
Chance magazine, (E): 640
 Concerts, (E): 261, 262
Consumer Reports, (BE): 5, 7, (E): 15, 407, 576–577, 675
 E-books, (E): 156
The Economist, (IE): 5
Journal of Applied Psychology, (E): 580–581
Journal of the American Medical Association, (E): 575, (IE): 487
Literary Digest, (IE): 321, 334
 Motion Picture Association of America (MPAA), (E): 51
 Movies, (E): 51–52, 55, 56, 59, 86, 90, 162, 264, 299–300
 Music, (E): 120, 148, 155, 160, 261, 262, 364, 402, 583, 603, (IE): 492
New England Journal of Medicine, (E): 637, 638
The New York Times, (BE): 332, (IE): 104
 News Reporting, (BE): 423, 425, (E): 13, 55, 87, 113, 534, 573, (JC): 425
 Online Magazines, (E): 503
 Public Opinion Surveys, (IE): 320
Readers' Digest, (E): 370
 Rock Concerts, (E): 114, 150
 Social Networking, (BE): 6, 543–544, 545, 550, (E): 13, 403, 505, 506, 573, (IE): 2, 393, 423–424, (JC): 484, (SBS): 394
Sports Illustrated, (E): 223
The Wall Street Journal, (IE): 2
World Drug Report, (E): 674

Pharmaceuticals, Medicine, and Health

Adolescent Dangerous Behavior, (BE): 323, (IE): 323, (WCGW): 41–42
 Alcohol Consumption, (E): 364, 370–371, (IE): 395–396
 Alternative Medical Treatments, (BE): 353, (E): 339, 340, 363, 371
 Alternative Medicine, (E): 14–15, 160
American Journal of Health Behavior, (IE): 395
 Aspirin, (E): 500, (JC): 484
 Baldness, (E): 191
 Blindness, (E): 14
 Blood Pressure, (E): 91, 362, 364, 406, 408, (SBS): 173–174
 Blood Types, (E): 405
 Body Fat, (BE): 130, 653, 660–661, (E): 14, 53, 54, 59, 229–230, 304, 467, 468, 506, (IE): 276–278, 279–280, 282, 283, 286–287, 642–644, 645, 646, 655–656,

659, (RM): 418, 653–654, (SBS): 648–650, (WCGW): 295
 Body Mass Index (BMI), (BE): 416, (E): 506, (IE): 661–664
 Body Temperature, (E): 153, 465, 504, (RM): 488–489
 Brain, (E): 606, 679
 Caffeine, (E): 118
 Cancer, (BE): 177, (E): 85, 362, 368, 504, 538, 574, (IE): 177, (SBS): 71–73
 Causes of Death, (E): 56, 87
 Centers for Disease Control and Prevention, (BE): 416, (E): 56, 264, 574, (IE): 125, (WCGW): 41–42
 Cholesterol, (E): 118, 119, 152, 406, 408, 463, 506, 579, 584, 635, 676
 Color-Blindness, (E): 534
 Congenital Abnormalities, (E): 502–503
 Death Rates, (E): 365
 Diabetes, (BE): 507–508, 511, 512, 516, 518–519, 522, (IE): 661–664
 Dialysis, (E): 155
 Diet, (E): 85, 86, 88, 92, 362, 364, 369
 Diseases/illnesses/Injuries, (E): 262, 365, 439, 638–639
 Diseases/Illnesses/Injuries, (BE): 177, 507–508, 511, 512, 516, 518–519, 522, 555, (E): 57, 85, 92, 115, 191, 361, 362, 363, 364, 368, 370, 440, 504, 528, 530, 533, 538, 574, 575, 635, 638–639, (IE): 527, 623–624, 661–664, (JC): 558, (SBS): 71–73, 625–626, 627–628
 Drug Abuse, (E): 192, 227, 339, 361
 Drug Development, (E): 501
 Drug Use/Abuse, (E): 674
 Eating Disorders/Weight Issues, (E): 91, 92, 369, 502, 537, 574, 575, 603, 608–609
 Emotional Health, (E): 314, 362, 363, 365, 368, 575
 ESP, (E): 222
 Exercise, (E): 14, 91, 362, 363, 364
 Fertility, (E): 270
 Food and Drug Administration (FDA), (BE): 346, 508, 522
 Gene Therapy, (E): 368
 Genetics, (E): 312
 Gestation/Pregnancy/Childbirth, (BE): 448, 484, 611, 614, 622, (E): 14, 91, 154–155, 219, 228–229, 268, 308, 313, 314, 362–363, 367, 403, 406, 407, 468, 469, 503, 533, 534, 536, 575, 634, 637, 640, (IE): 410–411, 419, 421, 441–442, 452–455, 482–483, (RM): 175–176, 325–326, 412, 422–423, (SBS): 451–452, 480–482, (WCGW): 430–431
 Harvard School of Public Health, (IE): 395
 Health Records, (E): 13
 Heart Disease, (BE): 511, 512, 516, 518–519, 522, (E): 57, 191, 361

Height and Weight, (BE): 126, 168–169, 416, (E): 155, 190–191, 304, 361, (IE): 125–127, 141, 169–171, 204, 246, 281, (JC): 131
 Hepatitis C, (IE): 623–624, (SBS): 625–626, 627–628
 HIV Testing, (E): 409
 Hospitals, (E): 93, 115, 365
 Insomnia, (E): 362, 363
Journal of the American Medical Association, (IE): 487, 511
 Life Expectancy, (E): 54, 191, 263, 270, 274, 275, 318, (IE): 243–244
 Mammograms, (E): 574
 Manual Dexterity, (IE): 585, 587, (RM): 596–597, (SBS): 593–594
 Marijuana, (E): 61
 Mayo Clinic, (IE): 410
 Medical Testing, (E): 528
 Medical Treatments, (JC): 354, 366
 Medication Side Effects, (E): 501
 Medication Trials, (BE): 507–508, 511, 512, 518–519, 522, (E): 13, 92, 190, 368, 370, 500, 502, 574, (IE): 527
 Memory, (E): 579–580, 583, 584
 Menopause, (E): 362
 National Center for Biotechnology, (BE): 345
 National Center for Chronic Disease Prevention and Health Promotion, (BE): 323, (IE): 323
 National Center for Health Statistics (NCHS), (E): 13, (IE): 410, 441
 National Health and Nutrition Examination Survey (NHANES), (BE): 125, (IE): 125
 National Institutes of Health, (BE): 345, (IE): 661
New England Journal of Medicine (NEJM), (BE): 507, 511, 516, 518, (E): 637, 638
 Omega-3, (E): 363
 Pain, (E): 574
 Pulse Rates, (E): 578
 Sleep, (IE): 484–485, 547–548, (RM): 73–74, (SBS): 485–486, 548–549
 Sleep Foundation, (SBS): 485
 Smoking, (E): 88, 119, 264–265, 313, 315, 368, 370, 501, 534, 536, 574, 575, (IE): 107–108, 177
 Stress Testing, (E): 13
 Tattoos, (E): 92, (IE): 623–624, 627–628, (SBS): 625–626
 TB Screening Test, (SBS): 397–398
 Therapeutic Touch (TT), (IE): 511–512
 Treatment Modalities, (E): 57
 University of Texas Southwestern Medical Center, (IE): 623
 Vaccinations, (E): 530, 574
 Veterinary Medicine, (E): 528
 Vision, (E): 464–465
 Vitamins, (E): 362, 365, 368, 439
 Women's Health Initiative, (BE): 345

Politics and Popular Culture

Approval Ratings, (E): 527
 Arrests, (E): 640
 Convention Bounce, (E): 576
 Crime and Television Watching, (E): 189
 Election Polls, (E): 534, 535, 573, 575, 576
 Embrace vs. Protect Attitudes, (E): 684
 Liberalism/Conservatism, (E): 89, 90, 92, 371, 635
 Petitions, (E): 537–538
 Pets, (E): 403, 536, 574
 Political Parties, (E): 85, 86, 161, 408, 639
 Roller Coasters, (BE): 99–100, 105, (E): 190, 193, 194, 222, 223, 340, (IE): 288–291, 657
 Socks, (E): 539
 Statue of Liberty, (E): 158
 Television Violence, (E): 580–581
Titanic Sinking, (BE): 67–68, 78–79, (E): 162, 403, 638, 640, (IE): 17–19, 24, 27, 31–35, 36, 75–76, (RM): 22–23, (WCGW): 41, 43, 81
 Voting and Elections, (E): 13, 14, 339, 340, 371, 407, 534, 535, 573, 575, 576, (IE): 240–241, 320, 321, 334
 Zodiac Signs, (IE): 610, (SBS): 614–616

Quality Control

Airplanes, (IE): 492
 Assembly Line, (E): 529
 Bottling, (E): 368
 Brewing, (IE): 445–446
 Cars, (E): 533
 Catheters, (E): 530
 Chips, (E): 530
 Fireworks, (E): 368
 Food Inspection and Safety, (BE): 483–484, (E): 341
 Light Bulbs, (E): 406
 Oranges, (E): 368
 Pet Food, (BE): 347, 352
 Product Inspections and Testing, (E): 471, 503, 533, 536–537, 539
 Product Ratings and Evaluations, (E): 15, 576–577, (JC): 37
 Recalls, (E): 539
 Underground Storage Tanks, (E): 533

Real Estate

Home Ownership Rate, (E): 528
 Home Sales, (BE): 278
 Housing Prices, (BE): 280, (E): 191, 193–194, 222, 223, 224, 299, 301, 302, 303, 317, 462–463, 468, 578, 675, 683–684, (JC): 205–206, 209, (SBS): 284–286
 Mortgages, (E): 191, 224, 465
 Racial Discrimination, (E): 640

Salary and Benefits

CEO Compensation, (BE): 130–131, (E): 151–152, 464, 467, 468, (IE): 105–106, 443–444, (WCGW): 142–143
 Earnings of College Graduates, (E): 672, 673, 674
 Employee Benefits, (E): 407
 Income and Housing Costs, (E): 191
 Salaries/Payroll, (BE): 130–131, (E): 58, 90, 147, 148, 151–152, 161, 188, 224, 273, 304, 463, (IE): 105–106, (WCGW): 142–143

Sales and Retail

Appliances, (E): 301
 Assets and Sales, (E): 316–317
 Books, (E): 187, 220, 221
 Coffee, (E): 189
 Customer Service, (E): 147
 Diamonds, (BE): 292–293
 Discounts, (E): 404
 Groceries, (E): 13, 462, 527
 Online Shopping, (E): 13, 188, 225
 Profits, (E): 680
 Purchase Amounts, (E): 572–573
 Sales, (E): 61
 Sales Promotions, (E): 406
 Shopping, (E): 403
 Socks, (BE): 474, (IE): 475–476

Science

Abalones, (E): 262
 Alligators, (E): 231
 Arm Length, (E): 341
 Astronomy, (E): 113, 195, 274
 Birds, (E): 57, 60–61, 222, 464, 527, (JC): 250
 Cattle, (E): 147, 149, 150, 152
 Cigarettes, (E): 223, 224
 Craters, (BE): 647–648, 655
 Crocodiles, (E): 312
 Deer Ticks, (E): 440
 Dexterity Testing, (E): 188, 305, (IE): 248–249
 Diet, (SBS): 71–73
 Dowsing, (E): 365, 502
 Draining Tanks, (E): 316
 Eggs, (E): 153
 Elephants, (E): 269
 Eye Color, (E): 117
 Fish, (E): 341–342
 F-stops, (IE): 178–179, 181, (WCGW): 182
 Genetics, (E): 636, (IE): 618, (JC): 71
 Hair Color, (E): 117
 Hamsters, (E): 534
 Handedness, (E): 404, (SBS): 417–418
 Hippos, (E): 269
 Language, (E): 536

Laundry Detergent, (E): 365, 369
 Lumber, (E): 274
 Manatees, (E): 309, 310
 Mazes, (E): 470, 505–506
 Mouth Volume, (JC): 656
 Nail Polish, (IE): 352, 354, (SBS): 348–349
 Oranges, (E): 275
 Pendulums, (E): 273
 Penguins, (E): 318, (IE): 234, (SBS): 251–254
 Pi, (E): 636
 Pottery Glazes, (E): 368, 530
 Shoe Size and IQ, (IE): 242, (WCGW): 182–183
 Speed of Light, (E): 465–466
 Stigler, Stephen M., (E): 466
 Temperatures, (SBS): 100–101
 Time Judgments, (E): 370
 Trees, (E): 150, 222, 275, 312–313
 Vision, (E): 312
 Walking in Circles, (E): 15
 Weighing Bears, (E): 14
 Youthful Appearance, (E): 370

Service Industries and Social Issues

Adoption, (E): 54
 Athletics and Relationships, (BE): 382, 390
 Cantril Scale, (E): 435
 Charitable Solicitations, (E): 434–435, 438, 503, (JC): 547
 Crawling, (E): 464, 681
 Donors, (E): 151
 Families, (E): 308
 Fundraisers, (E): 539
 Gun Control, (E): 89
 Honesty, (BE): 559, (E): 14, 362
 Internet Surfing, (IE): 565
 Living Arrangements, (E): 535
 Marriage, (BE): 103, 238–239, (E): 52, 116, 263, 267, 268, 269, 435, 501, 535, 676, 677
 Men's Attitudes, (E): 576
 Occupy Wall Street Movement, (E): 538
 Online Dating, (IE): 64, 65–66, 78
 Paralyzed Veterans of America, (E): 438, 503
 Parental Activities, (E): 87–88
 Parenting, (E): 436, 573, 576
 Psychics, (E): 500
 Pubs, (E): 370–371
 Racial Discrimination, (BE): 624, 626–627, 628, (E): 636, 640
 Religion, (E): 536
 Safety at Play, (E): 536
 Sharing Personal Information, (E): 573
 Social Life, (E): 339
 Social Networking, (BE): 6, 543–544, 545, 550, (E): 13, 403, 505, 506, 573, (IE): 2, 393, 423–424, (JC): 484, (SBS): 394

Violence against Women, (E): 636
 Working Parents, (E): 639

Sports

Age of Players, (E): 573
 Archery, (E): 533
 Athlete Name Recognition, (E): 537
 Baseball, (BE): 611, 614, 622, (E): 60, 119–120, 159, 192–193, 223, 224, 273, 302, 339, 361, 371, 438, 580, 583, 608, 682, (SBS): 489–491
 Basketball, (E): 116, 403, 530
 Bowling, (E): 406
 Dog Racing, (E): 363
 Exercise, (E): 606
 Fans, (E): 528
 Fishing, (E): 535–536
 Football, (BE): 66, 69, 70, (E): 13, 58, 151, 231–232, 315, 362, 403, 503
 Golf, (E): 59, 466, 470, 501, 583
 Hockey, (E): 60, 672, 673, 674
 Indy 500, (E): 15, 16
 Injuries, (E): 57
 Kentucky Derby, (E): 15, 16, 118, 189, (IE): 26–27, 37, (WCGW): 44
 Motorcycle Riding, (BE): 244–245, 291, (IE): 507, (SBS): 509–510
 Olympics, (BE): 124, 126, 127, (E): 56–57, 114, 150, 159, 230, 314, (IE): 122–124, 594–595, (SBS): 589–590, (WCGW): 143
 Relationships, (BE): 382, 390
 Running, (E): 57, 301, 302, 305–306, 582, 605
 Skiing, (E): 114, 150, 469, 505
 Skydiving, (E): 366
 Soccer, (E): 407, 535
 Speed Skating, (E): 159, (IE): 594–595, (SBS): 589–590
Sports Illustrated, (E): 223
 Super Bowl, (BE): 66, 69, 70, (E): 13, 58, 362
 Swimming, (E): 269, 365, 581–582
 Tour de France, (E): 16, 270–271, (JC): 7
 Uniforms, (E): 407, 535
 Weightlifting, (E): 274, 275
 Wheelchair Marathons, (E): 609

Surveys and Opinion Polls

American Association for Public Opinion Research (AAPOR), (JC): 392
 American Community Survey, (JC): 450
 American Time Use Survey (ATUS), (BE): 25
 Approval Ratings, (E): 440
 Athlete Name Recognition, (E): 537
 Car Purchases, (E): 574

Cell Phones, (E): 402–403
Economist/YouGov, (E): 436
 Elections, (E): 534, 535, 573, 575, 576
 Gallup Polls, (BE): 20, 428, (E): 339, 408, 435, (IE): 319, 321, (SBS): 426–427
 GfK Roper, (E): 339, 406, 407
 International Bedroom Poll, (IE): 547
 Liberalism/Conservatism, (E): 371
Literary Digest, (IE): 334
 Margin of Error, (E): 437
New York Times, (BE): 332
 Pew Research Organization, (BE): 20–21, 322, 423, 425, (E): 52, 86, 89, 161, 363, 371, 405, 573, (JC): 421, (SBS): 383–384
 Phone Surveys, (E): 341
 Postpurchase Surveys, (E): 162
 Presidential Popularity, (E): 537
 Public Opinion Polls, (E): 155, 340, 405, 406, 407, 436, 437, 440, 573, (IE): 319, 320, (JC): 392
 Quinnipiac University Poll, (IE): 319
 Random Samples, (E): 340
 Smoking, (E): 436
 Student Surveys, (E): 63, 85, 88, 90, 162, 338, 339, 463
 Zip Code, (E): 191

Technology

Cell Phones, (BE): 76–78, (E): 191, 262, 341, 369, 402–403, 434, 437, 502, (JC): 421–422
 Chips, (E): 316
 Computers, (BE): 7, (E): 404, 533, 536–537, 680–681
 Device Usage, (BE): 20–21
 Disk Drives, (E): 187, 220
 E-Mail, (E): 53, 54, 528
 E-Readers, (E): 408
 Internet, (BE): 8, (E): 13, 55, 61, 113, 314, 372, 435, 436, 438, 573, (IE): 565
 MP3 Players, (E): 149
 National Strategy for Trusted Identities in Cyberspace, (BE): 6
 Online Dating, (IE): 64, 65–66, 78
 Online Magazines, (E): 503
 Social Networking, (BE): 6, 543–544, 545, 550, (E): 13, 403, 505, 506, 573, (IE): 2, 393, 423–424, (JC): 484, (SBS): 394
 Software, (E): 368, 529
 Stereograms, (E): 120–121, 577
 Telephone, (E): 5, 341
 Television, (E): 189, 263, 370, 404, 437, 530, 573, 580–581, (IE): 243–244, (SBS): 329–330
 Websites, (E): 434, 500, 527

Transportation

Accidents, (BE): 399, (E): 51, 53, 56, 87, 156–157, 158, 159, 438, 604, (IE): 395–396, 507
 Air Travel, (E): 56, 87, 111, 112–113, 195, 265, 271, 372, 408, 409, 466, 504, 527, (IE): 384–385, 492, (JC): 101, 121
 Bicycle Safety, (E): 14, 161–162
 Bridge Safety, (E): 228, 270, 369, (RM): 203–204, 644, (SBS): 202–204
 Bureau of Transportation Statistics, (JC): 101
 Car Manufacture, (E): 438, 635, 639
 Car Ownership, (E): 502
 Car Purchases, (E): 437, 574
 Car Repairs, (E): 404, 405, 501
 Car Speeds, (E): 149, 150, 267, (IE): 560
 Commute Distances, (BE): 588–589, 591–592
 Commute Times, (IE): 457, (JC): 131, (RM): 40–41, 133–134, (SBS): 455–456
 Distance Traveled, (E): 636
 Drivers' Licenses, (BE): 478, 479, 480, (E): 92
 Drivers Type, (E): 408
 Driving Speeds, (E): 404, 537, (RM): 102–103
 Emissions Testing, (E): 440, 528–529
 Engine Size, (E): 160–161
 Fuel Economy, (E): 15, 116–117, 119, 147, 191, 192, 221, 231, 271, 272, 275, 311–312, 341, 365, 371, 602, 606–607, 678, 679, 682–683, (IE): 141–142, 246–248, 281, (JC): 37, (SBS): 38–39
 Gas Prices, (E): 60, 116
 Gasoline Supply, (E): 315
 Horsepower, (E): 311, (IE): 281
 Motorcycles, (IE): 507
 Roadblocks, (E): 339
 Sobriety Checkpoints, (E): 408
 Speeding, (E): 404, 537
 Stopping Distances, (E): 273, 470, 607
 Stopping Times, (E): 262
 Texting While Driving, (IE): 2
 Tires, (E): 152–153
 Traffic/Parking/Safety, (BE): 385–386, (E): 13, 91, 190, 310, 465, 500, 527, 529, 539, 603–604, (IE): 373–375, 541, 542, (JC): 629, (SBS): 545–546
 Trains, (E): 403, 406
 Used Cars, (E): 226–227, 676–677
 Weights of Vehicles, (E): 194–195, 222, 316, 602

This page intentionally left blank

Stats Starts Here¹

WHERE ARE WE GOING?

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

But statistics *is* fun. That’s probably not what you heard on the street, but it’s true. Statistics is the science of learning from data. A little practice thinking statistically is all it takes to start seeing the world more clearly and accurately.

This is a book about understanding the world by using data. So we’d better start by understanding data. There’s more to that than you might have thought.



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

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

—Lewis Carroll,
Alice’s Adventures
in Wonderland

1.1 What Is Statistics?

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

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

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

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



FRAZZ © 2003 Jef Mallett. Distributed by Andrews McMeel Syndication. Reprinted with permission. All rights reserved.

“Data is king at Amazon. Clickstream and purchase data are the crown jewels at Amazon. They help us build features to personalize the Web site experience.”

—Ronny Kohavi,
former Director of Data
Mining and Personalization,
Amazon.com

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

Consider the following:

- ◆ If you have a Facebook account, you have probably noticed that the ads you see online tend to match your interests and activities. Coincidence? Hardly. According to *The Wall Street Journal* (10/18/2010),² much of your personal information has probably been sold to marketing or tracking companies. Why would Facebook give you a free account and let you upload as much as you want to its site? Because your data are valuable! Using your Facebook profile, a company might build a profile of your interests and activities: what movies and sports you like; your age, sex, education level, and hobbies; where you live; and, of course, who your friends are and what *they* like. From Facebook’s point of view, your data are a potential gold mine. Gold ore in the ground is neither very useful nor pretty. But with skill, it can be turned into something both beautiful and valuable. What we’re going to talk about in this book is how you can mine your own data and learn valuable insights about the world.
- ◆ 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 book, you’ll learn how to design and analyze experiments like this. You’ll learn how to interpret data and to communicate the message you see to others. You’ll also learn how to spot deficiencies and weaknesses in conclusions drawn by others that you see in newspapers and on the Internet every day. Statistics can help you become a more informed citizen by giving you the tools to understand, question, and interpret data.

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.”

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

³<http://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

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 2016, the company’s sales reached almost \$136 billion (more than 25% over the previous year). Amazon has sold a wide variety of merchandise, including a \$400,000 necklace, yak cheese from Tibet, and the largest book in the world. How did Amazon become so successful and how can it keep track of so many customers and such a wide variety of products? The answer to both questions is *data*.

But what are data? Think about it for a minute. What exactly *do* we mean by “data”? You might think that data have to be numbers, but data can be text, pictures, web pages, and even audio and video. If you can sense it, you can measure it. 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	B0000015Y6	440	B000002BK9	0.99
Canada	Detroit	440	105-1372500-0198646	N	B002MXA7Q0	Ohio	Y

Try to guess what they represent. Why is that hard? Because there is no *context*. If we don’t know what values are measured and what is measured about them, the values are meaningless. We can make the meaning clear if we organize the values into a **data table** such as this one:

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

Now we can see that these are purchase records for album download orders from Amazon. The column titles tell what has been recorded. Each row is about a particular purchase.

⁵<http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>

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.

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.”

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

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

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

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

How the Data Are Collected

How the data are collected can make the difference between insight and nonsense. As we’ll see later, data that come from a voluntary survey on the Internet are almost always worthless. One primary concern of statistics, to be discussed in Part III, is the design of sound methods for collecting data. Throughout this book, whenever we introduce data, we’ll provide a margin note listing the W’s (and H) of the data. Identifying the W’s is a habit we recommend.

The first step of any data analysis is to know what you are trying to accomplish and what you want to know. To help you use statistics to understand the world and make decisions, we’ll lead you through the entire process of *thinking* about the problem, *showing* what you’ve found, and *telling* others what you’ve learned. Every guided example in this book is broken into these three steps: *Think*, *Show*, and *Tell*. Identifying the problem and the *who* and *what* of the data is a key part of the *Think* step of any analysis. Make sure you know these before you proceed to *Show* or *Tell* anything about the data.



EXAMPLE 1.1

Identifying the *Who*

In 2015, *Consumer Reports* published an evaluation of 126 tablets from a variety of manufacturers.

QUESTION: Describe the population of interest, the sample, and the *Who* of the study.

ANSWER: The magazine is interested in the performance of tablets currently offered for sale. It tested a sample of 126 tablets, which are the *Who* for these data. Each tablet selected represents all similar tablets offered by that manufacturer.

1.3 Variables

The characteristics recorded about each individual are called **variables**. They are usually found as the columns of a data table with a name in the header that identifies what has been recorded. In the Amazon data table we find the variables *Order Number*, *Name*, *State/Country*, *Price*, and so on.

Categorical Variables

Some variables just tell us what group or category each individual belongs to. Are you male or female? Pierced or not? We call variables like these **categorical**, or **qualitative variables**. (You may also see them called **nominal variables** because they name categories.) Some variables are clearly categorical, like the variable *State/Country*. Its values are text and those values tell us what category the particular case falls into. But numerals are often used to label categories, so categorical variable values can also be numerals. For example, Amazon collects telephone area codes that *categorize* each phone number into a geographical region. So area code is considered a categorical variable even though it has numeric values. (But see the story in the following box.)



“Far too many scientists have only a shaky grasp of the statistical techniques they are using. They employ them as an amateur chef employs a cookbook, believing the recipes will work without understanding why. A more *cordon bleu* attitude . . . might lead to fewer statistical soufflés failing to rise.”

—The Economist, June 3, 2004, “Sloppy stats shame science”

AREA CODES—NUMBERS OR CATEGORIES?

The *What* and *Why* of area codes are not as simple as they may first seem. When area codes were first introduced, AT&T was still the source of all telephone equipment, and phones had dials.

To reduce wear and tear on the dials, the area codes with the lowest digits (for which the dial would have to spin least) were assigned to the most populous regions—those with the most phone numbers and thus the area codes most likely to be dialed. New York City was assigned 212, Chicago 312, and Los Angeles 213, but rural upstate New York was given 607, Joliet was 815, and San Diego 619. For that reason, at one time the numerical value of an area code could be used to guess something about the population of its region. Since the advent of push-button phones, area codes have finally become just categories.

Descriptive responses to questions are often categories. For example, the responses to the questions “Who is your cell phone provider?” and “What is your marital status?” yield categorical values. When Amazon considers a special offer of free shipping to customers, it might first analyze how purchases have been shipped in the recent past. Amazon might start by counting the number of purchases shipped in each category: ground transportation, second-day air, and next-day air. Counting is a natural way to summarize a categorical variable such as *Shipping Method*. Chapter 2 discusses summaries and displays of categorical variables more fully.

Quantitative Variables

When a variable contains measured numerical values with measurement *units*, we call it a **quantitative variable**. Quantitative variables typically record an amount or degree of something. For quantitative variables, its measurement **units** provide a meaning for the numbers. Even more important, units such as yen, cubits, carats, angstroms, nanoseconds, miles per hour, or degrees Celsius tell us the *scale* of measurement, so we know how far apart two values are. Without units, the values of a measured variable have no meaning. It does little good to be promised a raise of 5000 a year if you don't know whether it will be paid in Euros, dollars, pennies, yen, or Mauritanian Ouguiya (MRO).⁶ We'll see how to display and summarize quantitative variables in Chapter 2.

Sometimes a variable with numeric values can be treated as either categorical or quantitative depending on what we want to know from it. Amazon could record your *Age* in years. That seems quantitative, and it would be if the company wanted to know the average age of those customers who visit their site after 3 AM. But suppose Amazon wants to decide which album to feature on its site when you visit. Then thinking of your age in one of the categories Child, Teen, Adult, or Senior might be more useful. So, sometimes whether a variable is treated as categorical or quantitative is more about the question we want to ask rather than an intrinsic property of the variable itself.

Identifiers

For a categorical variable like *Survived*, each individual is assigned one of two possible values, say *Alive* or *Dead*.⁷ But for a variable with ID numbers, such as a *student ID*, each individual receives a unique value. We call a variable like this, which has exactly as many values as cases, an **identifier variable**. Identifiers are useful, but not typically for analysis.

Amazon wants to know who you are when you sign in again and doesn't want to confuse you with some other customer. So it assigns you a unique identifier. Amazon also wants to send you the right product, so it assigns a unique Amazon Standard Identification Number (ASIN) to each item it carries. You'll want to recognize when a variable is playing the role of an identifier so you aren't tempted to analyze it.

Identifier variables themselves don't tell us anything useful about their categories because we know there is exactly one individual in each. Identifiers are part of what's called **metadata**, or data about the data. Metadata are crucial in this era of large data sets because by uniquely identifying the cases, they make it possible to combine data from different sources, protect (or violate) privacy, and provide unique labels.⁸ Many large databases are *relational* databases. In a relational database, different data tables link to one another by matching identifiers. In the Amazon example, the *Customer Number*, *ASIN*, and *Transaction Number* are all identifiers. The IP (Internet Protocol) address of your computer is another identifier, needed so that the electronic messages sent to you can find you.

Ordinal Variables

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

But what are the units? There is certainly an *order* of perceived worth: Higher numbers indicate higher perceived worth. A course that averages 4.5 seems more valuable than one that averages 2, but we should be careful about treating *Educational Value* as purely

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.

⁶As of 10/26/2016 \$1 = 357.95 MRO

⁷Well, maybe three values if you include Zombies.

⁸The National Security Agency (NSA) made the term "metadata" famous in 2014 by insisting that they only collected metadata on US citizens phone calls and text messages, not the calls and messages themselves. They later admitted to the bulk collection of actual data.

quantitative. To treat it as quantitative, she'll have to imagine that it has “educational value units” or some similar arbitrary construct. Because there are no natural units, she should be cautious. Variables that report order without natural units are often called **ordinal variables**. But saying “that’s an ordinal variable” doesn’t get you off the hook. You must still look to the *Why* of your study and understand what you want to learn from the variable to decide whether to treat it as categorical or quantitative.

EXAMPLE 1.2

Identifying the *What* and *Why* of Tablets

RECAP: A *Consumer Reports* article about 126 tablets lists each tablet’s manufacturer, price, battery life (hrs.), the operating system (Android, iOS, or Windows), an overall quality score (0–100), and whether or not it has a memory card reader.

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

ANSWER: The variables are

- manufacturer (categorical)
- price (quantitative, \$)
- battery life (quantitative, hrs.)
- operating system (categorical)
- quality score (quantitative, no units)
- memory card reader (categorical)

The magazine hopes to provide consumers with the information to choose a good tablet.



JUST CHECKING

In the 2004 Tour de France, Lance Armstrong made history by winning the race for an unprecedented sixth time. In 2005, he became the only 7-time winner and set a new record for the fastest average speed—41.65 kilometers per hour—that stands to this day. In 2012, he was banned for life for doping offenses, stripped of all of his titles and his records expunged. You can find data on all the Tour de France races in the data set **Tour de France 2016**. Here are the first three and last seven lines of the data set. Keep in mind that the entire data set has over 100 entries.

1. List as many of the W’s as you can for this data set.
2. Classify each variable as categorical or quantitative; if quantitative, identify the units.

Year	Winner	Country of Origin	Age	Team	Total Time (h/min/s)	Avg. Speed (km/h)	Stages	Total Distance Ridden (km)	Starting Riders	Finishing Riders
1903	Maurice Garin	France	32	La Française	94.33.00	25.7	6	2428	60	21
1904	Henri Cornet	France	20	Cycles JC	96.05.00	25.3	6	2428	88	23
1905	Louis Trousseller	France	24	Peugeot	112.18.09	27.1	11	2994	60	24
...										
2010	Andy Schleck	Luxembourg	25	Saxo Bank	91.59.27	39.59	20	3642	180	170
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	Christopher Froome	Great Britain	28	Sky	83.56.40	40.55	21	3404	198	169
2014	Vincenzo Nibali	Italy	29	Astana	89.56.06	40.74	21	3663.5	198	164
2015	Christopher Froome	Great Britain	30	Sky	84.46.14	39.64	21	3660.3	198	160
2016	Christopher Froome	Great Britain	31	Sky	89.04.48	39.62	21	3529	198	174



THERE'S A WORLD OF DATA ON THE INTERNET

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

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

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

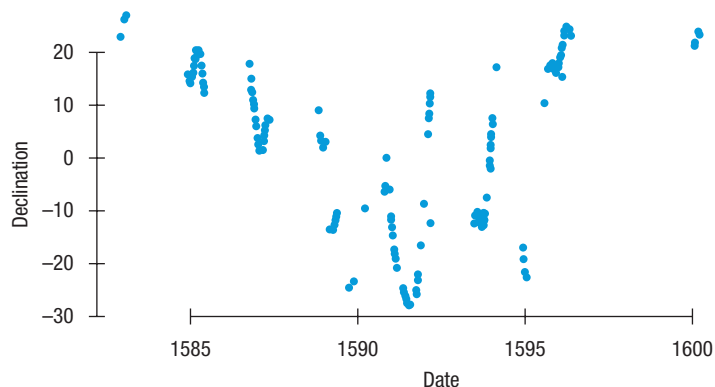
1.4 Models

What is a **model** for data? Models are summaries and simplifications of data that help our understanding in many ways. We'll encounter all sorts of models throughout the book. A model is a simplification of reality that gives us information that we can learn from and use, even though it doesn't represent reality exactly. A model of an airplane in a wind tunnel can give insights about the aerodynamics and flight performance of the plane even though it doesn't show every rivet.⁹ In fact, it's precisely because a model is a simplification that we learn from it. Without making models for how data vary, we'd be limited to reporting only what the data we have at hand says. To have an impact on science and society we'll have to generalize those findings to the world at large.

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

Figure 1.1

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

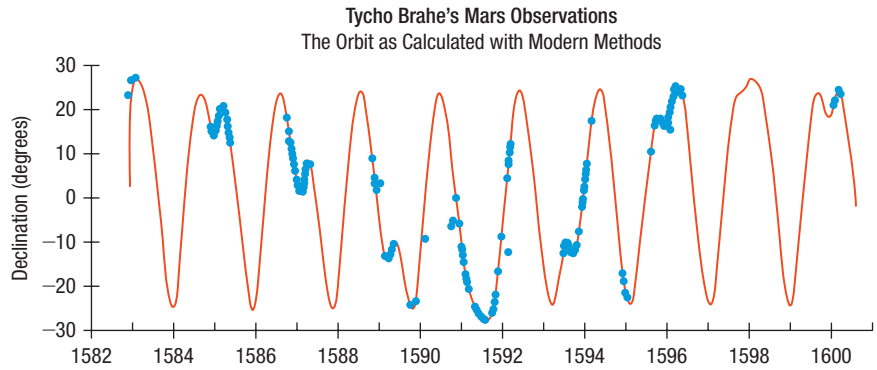


Here, using modern statistical methods is a plot of the model predictions from the data:

⁹Or tell you what movies you might see on the flight.

Figure 1.2

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



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

WHAT CAN GO WRONG?

- ◆ **Don't label a variable as categorical or quantitative without thinking about the data and what they represent.** The same variable can sometimes take on different roles.
- ◆ **Don't assume that a variable is quantitative just because its values are numbers.** Categories are often given numerical labels. Don't let that fool you into thinking they have quantitative meaning. Look at the context.
- ◆ **Always be skeptical.** One reason to analyze data is to discover the truth. Even when you are told a context for the data, it may turn out that the truth is a bit (or even a lot) different. The context colors our interpretation of the data, so those who want to influence what you think may slant the context. A survey that seems to be about all students may in fact report just the opinions of those who visited a fan website. The question that respondents answered may be posed in a way that influences responses.

CHAPTER REVIEW



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

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

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