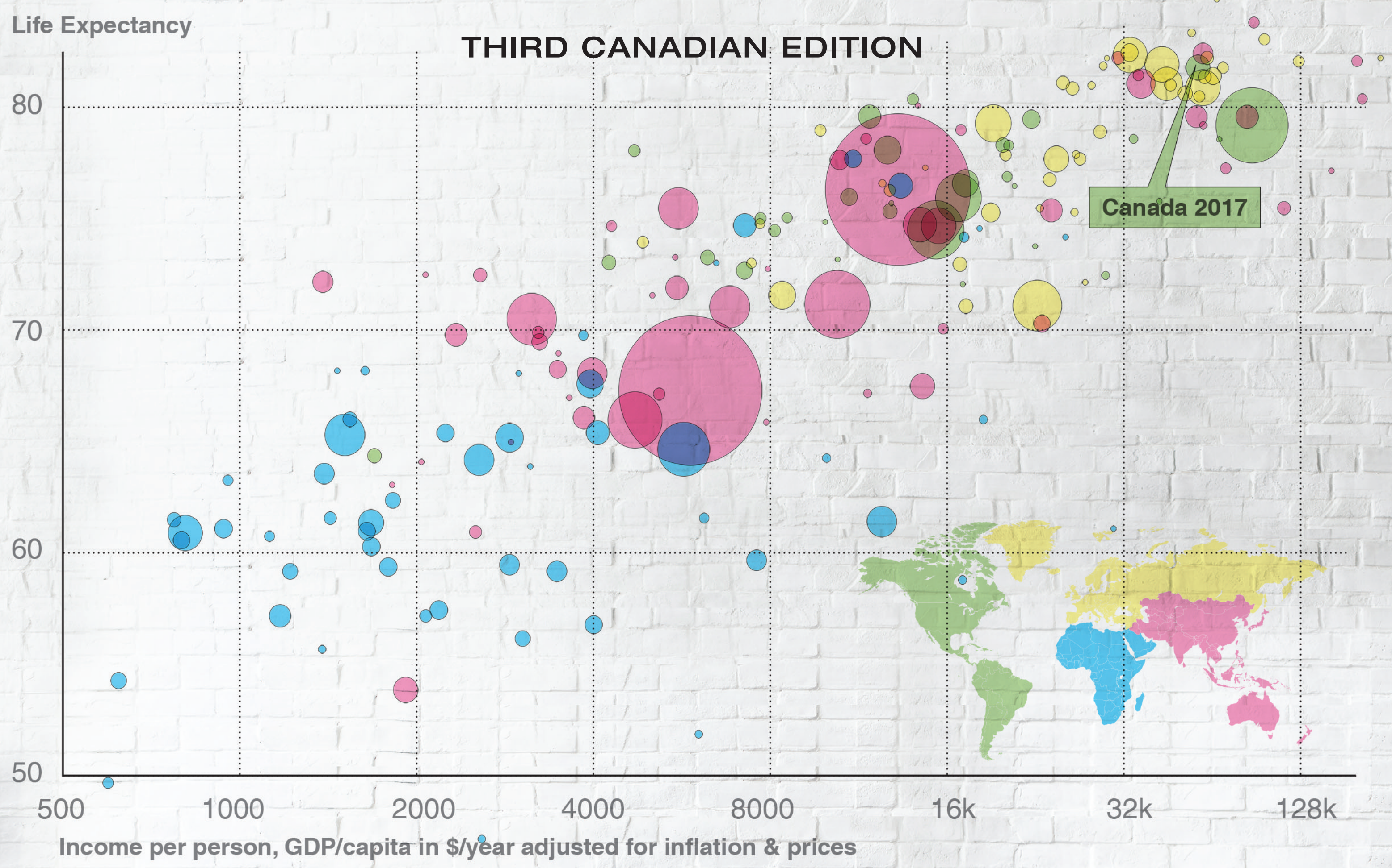
DE VEAUX • VELLEMAN • BOCK • VUKOV • WONG

STATS

DATA AND MODELS

THIRD CANADIAN EDITION



This page intentionally left blank

Selected Formulas

$Range = Max - Min$

$IQR = Q3 - Q1$

Outlier Rule-of-Thumb: $y < Q1 - 1.5 \times IQR$ or $y > Q3 + 1.5 \times IQR$

$$\bar{y} = \frac{\sum y}{n}$$

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$$

$$z = \frac{y - \mu}{\sigma} \text{ (model based)}$$

$$z = \frac{y - \bar{y}}{s} \text{ (data based)}$$

$$r = \frac{\sum z_x z_y}{n - 1}$$

$$\hat{y} = b_0 + b_1 x \qquad \text{where } b_1 = r\frac{s_y}{s_x} \text{ and } b_0 = \bar{y} - b_1\bar{x}$$

$P(\mathbf{A}) = 1 - P(\mathbf{A}^C)$

$P(\mathbf{A} \text{ or } \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \text{ and } \mathbf{B})$

$P(\mathbf{A} \text{ and } \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B} | \mathbf{A})$

$$P(\mathbf{B} | \mathbf{A}) = \frac{P(\mathbf{A} \text{ and } \mathbf{B})}{P(\mathbf{A})}$$

A and **B** are independent if $P(\mathbf{B} | \mathbf{A}) = P(\mathbf{B})$. Then $P(\mathbf{A} \text{ and } \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B})$

$E(X) = \mu = \sum x P(x)$	$Var(X) = \sigma^2 = \sum (x - \mu)^2 P(x)$
$E(X \pm c) = E(X) \pm c$	$Var(X \pm c) = Var(X)$
$E(aX) = aE(X)$	$Var(aX) = a^2 Var(X)$
$E(X \pm Y) = E(X) \pm E(Y)$	$Var(X \pm Y) = Var(X) + Var(Y)$ if X and Y are independent

Binomial: $P(x) = {}_nC_x p^x q^{n-x}$ $\quad \mu = np \quad \sigma = \sqrt{npq}$

$$\hat{p} = \frac{x}{n} \qquad \mu(\hat{p}) = p \qquad SD(\hat{p}) = \sqrt{\frac{pq}{n}}$$

Sampling distribution of $\bar{y}$:

(CLT) As n grows, the sampling distribution approaches the Normal model with

$$\mu(\bar{y}) = \mu \qquad SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$$

Inference:

Confidence interval for Parameter = *Estimate* ± *Critical value* × *SE*(*Estimator*)

Test statistic $= \dfrac{Estimate - Parameter}{SE(Estimator)}$ [Replace *SE* by *SD* if latter is known]

Parameter	Estimator	***SD*** **(Estimator)**	***SE*** **(Estimator)**
p	$\hat{p}$	$\sqrt{\frac{pq}{n}}$	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$
μ	$\bar{y}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$
$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
μ_d	$\bar{d}$	$\frac{\sigma_d}{\sqrt{n}}$	$\frac{s_d}{\sqrt{n}}$
σ_ε	$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$	(divide by $n-k-1$ in multiple regression)	
β_1	b_1	(in simple regression)	$\frac{s_e}{s_x\sqrt{n-1}}$
μ_ν	$\hat{y}_\nu$	(in simple regression)	$\sqrt{SE^2(b_1)\cdot(x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}$
y_ν	$\hat{y}_\nu$	(in simple regression)	$\sqrt{SE^2(b_1)\cdot(x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$

Pooling: For testing difference between proportions: $\hat{p}_{pooled} = \dfrac{y_1 + y_2}{n_1 + n_2}$

For testing difference between means (when $\sigma_1 = \sigma_2$): $s_p = \sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

Substitute these pooled estimates in the respective SE formulas for both groups when assumptions and conditions are met.

Chi-square: $\chi^2 = \sum \dfrac{(Obs - Exp)^2}{Exp}$

One-way ANOVA: $SS_T = \sum\sum(\bar{y}_j - \bar{\bar{y}})^2$; $\quad MS_T = SS_T/(k - 1)$

$SS_E = \sum\sum(\bar{y}_{ij} - \bar{y}_j)^2$; $\quad MS_E = SS_E/(N - k)$

$F = MS_T/MS_E$ with df $= (k - 1, N - k)$

This page intentionally left blank

STATS

DATA AND MODELS

This page intentionally left blank

THIRD CANADIAN EDITION

STATS

DATA AND MODELS

RICHARD D. DE VEAUX
WILLIAMS COLLEGE

PAUL F. VELLEMAN
CORNELL UNIVERSITY

DAVID E. BOCK
CORNELL UNIVERSITY

AUGUSTIN M. VUKOV
UNIVERSITY OF TORONTO

AUGUSTINE C.M. WONG
YORK UNIVERSITY

Pearson

Vice President, Editorial: Anne Williams
Acquisitions Editor: Mark Grzeskowiak
Marketing Manager: Euan White
Content Manager: Emily Dill
Project Manager: Pippa Kennard
Content Developer: Paul Donnelly
Media Developer: Kelli Cadet
Production Services: Cenveo® Publisher Services
Permissions Project Manager: Joanne Tang
Photo Permissions Research: Integra Publishing Services
Text Permissions Research: Integra Publishing Services
Interior Designer: Anthony Leung
Cover Designer: Anthony Leung
Cover Image: Rawpixel.com/Shutterstock.com, Graph is based on free material from GAPMINDER.ORG, CC-BY LICENSE
Vice-President, Cross Media and Publishing Services: Gary Bennett

Pearson Canada Inc., 26 Prince Andrew Place, North York, Ontario M3C 2H4.

9780134301051

10 9 8 7 6 5 4 3 2 1

Library and Archives Canada Cataloguing in Publication

De Veaux, Richard D., author
Stats : data and models / Richard D. De Veaux, Williams College, Paul F. Velleman, Cornell University, David E. Bock, Cornell University, Augustin M. Vukov, University of Toronto, Augustine C.M. Wong, York University.— Third Canadian edition.

Includes index.

ISBN 978-0-13-430105-1 (hardcover)

1. Statistics—Textbooks. 2. Mathematical statistics—Textbooks. I. Velleman, Paul F., 1949-, author II. Bock, David E., author III. Vukov, Augustin M., author IV. Wong, Augustine C. M., author V. Title.

QA276.12.D42 2017 519.5 C2017-902493-0

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

To my adopted country, Canada, its amazing citizens,
indigenous peoples, immigrants and refugees, whose stories inspire me,
and the University of Toronto without which this
dedication would never have happened.

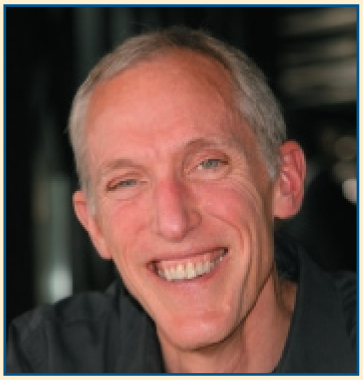
—Gus

To Camilla and Octavia for their
understanding and encouragement

—Augustine

This page intentionally left blank

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." Since 1994, he has been Professor of Statistics at Williams College. 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). 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 25 years he has consulted for such Fortune 500 companies as American Express, Hewlett-Packard, Alcoa, DuPont, Pillsbury, General Electric, and Chemical Bank. Because he consulted with Mickey Hart on his book *Planet Drum*, he has also sometimes been called the "Official Statistician for the Grateful Dead." His real-world experiences and anecdotes illustrate many of this book's chapters.

Dick holds degrees from Princeton University in Civil Engineering (B.S.E.) and Mathematics (A.B.) and from Stanford University in Dance Education (M.A.) and Statistics (Ph.D.), where he studied dance with Inga Weiss and Statistics with Persi Diaconis. 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. 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) (lib.stat.cmu.edu/DASL/), which provides data sets for teaching Statistics. Paul's understanding of using and teaching with technology informs much of this book's approach.

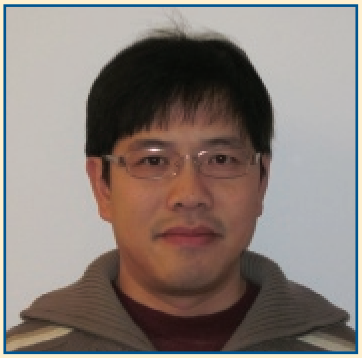
Paul has taught Statistics at Cornell University since 1975. He holds an A.B. from Dartmouth College in Mathematics and Social Science, and M.S. and Ph.D. degrees in Statistics from Princeton University, where he studied with John Tukey. His research often deals with statistical graphics and data analysis 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 recently 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.

Augustin M. Vukov has been teaching at the University of Toronto for several decades. For much of that time, he was the course coordinator for *The Practice of Statistics*—a large multi-section course designed to introduce the basic concepts and practice of Statistical Science to students from a great variety of (mostly non-mathematical) disciplines. Having taught so many students their required Stats course, he was not surprised to hear his new family doctor, his new ophthalmologist and a prospective new dentist say “Your name looks familiar. Ah yes, I used to sit in your Statistics lecture!”. A big fan of Minitab™ software for use in introductory Statistics courses, he has also authored several Minitab manuals.



Augustine C.M. Wong is a professor of Statistics at York University. He completed his Ph.D. at the University of Toronto in 1990 and taught at the University of Waterloo and University of Alberta before coming to York in 1993. His research interests include asymptotic inference, computational methods in statistics, and likelihood-based methods. He is an author or co-author of over 80 research articles and 2 book chapters. At York, he teaches various statistics courses at both the undergraduate and graduate levels.

Contents

*Optional section.

Part III Gathering Data

Part IV Randomness and Probability

Part V From the Data at Hand to the World at Large

*Optional section.

*Optional section.

Part VII Modelling the World at Large

Part VIII Distribution-free Methods

*Optional section.

Preface

We are often asked why we write Statistics texts. After all, it takes a lot of work to find new and better examples, to keep datasets current, and to make a book an enjoyable and effective learning tool. So we thought we'd address that question first.

We do it because it's fun.

Of course, we care about teaching students to think statistically; we are teachers and professional statisticians. And we want to change the way students think about the world. But Statistics can be a particularly challenging subject to teach. The student encounters many new concepts, many new methods, and many new terms. Our challenge has been to write a book that students would read, learn from, and enjoy. And we recommit to this goal with each new edition.

The book you hold is more succinct and, we hope, even more readable than previous editions. Of course, we've kept our conversational style and background anecdotes.[1] But we've tightened discussions and adjusted the order of some topics. The story we tell about Statistics moves even more quickly to interesting real-world questions.

Increasingly teachers are using examples from Statistics to provide intuitive examples of how a little bit of math can help us say a lot about the world. Students expect Statistics to be about real-world insights. This edition of *Stats: Data and Models* keeps students engaged and interested because we show Statistics in action right from the start. Students will be solving problems of the kind they're likely to encounter in real life sooner. In Chapter 4, they will be comparing groups and in Chapter 6, they'll see relationships between two quantitative variables—and, of course, always with real, modern data. By emphasizing statistical thinking, these early chapters lay the foundation for the more advanced chapters on Analysis of Variance and Multiple Regression at the end of the book, telling a consistent story throughout.

We hope our book will help students realize Statistics can be fun, useful, and empowering. There are few things more rewarding than discovering the world anew through the lens of statistical analysis. And few things more fun for authors than helping students make those discoveries.

So, What's New in This Edition?

We've scrutinized literally every sentence and paragraph throughout the book, and rewritten many sections to make them clearer and more interesting, along with introducing some new up-to-the-minute motivating examples.

We've added some new features, each with the goal of making it even easier for students to put the concepts of Statistics together into a coherent whole.

1. New and improved pedagogical tools: We've replaced the list of sections at the start of each chapter with a more informative list of *Learning Objectives*. We've also replaced *Where Are We Going* with *Connections*, while deleting the similarly named item that used to sit (we believe, almost invisibly) at the end of chapters in earlier editions. Chapter study materials at the end of chapters still include *What Have We Learned* including the *Review of Terms*. Students who understand the latter items are well prepared for future chapters (not to mention well on their way to being ready for any tests).

2. Streamlined content: Our reorganization has shortened the book from 30 to 29 chapters (Chapter 29 is only available as part of the eText in MyLab™ Statistics). We've also grouped some important concepts, in new and better presentation orders. Each chapter is still a focused discussion, and most can be taught in one lesson.

[1] and our footnotes.

3. Content changes: Here's how we've reorganized or changed the content:

   a. We've reorganized material in *Chapter 3*: *Displaying and Summarizing Quantitative Data*. Instead of measures of centre followed by measures of spread, we start with the more intuitive measures: median, IQR, and 5-number summary, followed later by the mean and standard deviation.

   b. We've removed the chapter *Understanding Randomness*, while moving its section on Random Numbers into the following chapter on Sample Surveys.

   c. We've tried to improve the organization of material in Parts V–VII:

   - *Inference About Means* appears earlier now, in Part V. We still lead the discussion of inference with inference for proportions (for reasons we explain in the *Test Bank and Resource Guide*), but now we turn sooner to inference for means so students can see the methods side by side. Students can then also see that the reasoning is essentially the same.
   - *Comparing Two Proportions* was moved from Part V to Part VI, so that it now follows the similar discussion for comparing two means, with the chapter on chi-square tests immediately after.
   - *Inference for Regression* was moved from Part VII to Part VI, where it more logically belongs, though in the same chapter order as before.

   d. We've added an optional section on the Regression ANOVA to the *Inference for Regression* chapter.

4. Section Exercises: A new feature of this edition is the inclusion of *Section Exercises.* In addition to the usual end-of-chapter exercises, at the end of each section, you will now find these new single-concept exercises.

5. Exercises: We've updated most exercises that use real-world data, retired some that were getting old, and added new exercises. As in previous editions, many of the exercises feature Canadian content and data with up-to-date references to Indigenous peoples, sports, health care, education, the environment, and other social and political issues.

Our Approach

Statistical Science is practiced with technology. We think a modern Statistics text should recognize that fact from the start. And so do our students. You won't find tedious calculations worked by hand. But you will find equation forms that favour intuition over calculation. You'll find extensive use of real data—even large data sets. And, most important, 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 *find* the answer?" but "how do you *think* about the answer?"

We've been guided in the choice and order of topics by several fundamental principles. First, we have tried to ensure that each new topic fits into the growing structure of understanding that we hope students will build. We know that learning a new subject like Statistics requires putting together many new ideas, and we have tried to make that process as natural as possible. That goal has led us to cover some topics in a different order than is found in other texts.

For example, we introduce inference by looking at a confidence interval for a proportion rather than the more traditional approach that starts with the one-sample mean. Everyone has seen opinion polls. Most people understand that pollsters use a sample of voters to try to predict the preferences of the entire population and that the result is an estimate with a margin of error. Showing how to construct and interpret a confidence interval in this context introduces key concepts of inference in a familiar setting.

We next examine hypothesis tests for a proportion, which use formulas and a test statistic that is now familiar, so it is easier to focus on the logic and structure of a formal hypothesis test. When we then turn our attention to the topic of small-sample inference for means, the only new issue is the t-distribution. However, should you wish to discuss the z-test and confidence interval for means (known sigma or large n) earlier, in tandem with the z-procedures for proportions, optional sections are included in those chapters.

Textbooks can be defined more by what they choose not to cover than by what they do cover. We've structured this text so that each new topic fits into a student's growing understanding. Several topic orders can support this goal. While this text contains material sufficient for two semesters, it offers good choices for those teaching a one-semester course, for which we recommend coverage of Chapters 1–21, followed by (time permitting) one or more of Chapters 22, 23, 24, and 28. Some topics in Chapters 9–10 on sampling and experimental design may be de-emphasized or omitted entirely depending on the needs of the particular audience. Likewise, the material on probability and random variables in Chapters 11–13 is considerable and portions may be de-emphasized. Optional sections, indicated by asterisks, are indeed just that and may be omitted without harm. We have also differentiated the coverage in our Canadian edition from its U.S. counterpart in some ways. We added two additional chapters (authored by A. Vukov), on nonparametric procedures (Chapter 28) and bootstrapping (Chapter 29). We also consistently put more emphasis on power & sample size issues for tests of hypothesis.

GAISE Guidelines

The Guidelines for Assessment and Instruction in Statistics Education (GAISE) report adopted by the American Statistical Association urges that Statistics education should

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

We've designed our text and supporting materials to encourage this approach to the introductory course. We urge you to think about these guidelines with each class meeting.

Our Goal: Read This Book!

The best text in the world is of little value if students don't read it. Here are some of the ways we have made this edition even more approachable:

- ***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. We've tightened the discussions and removed digressions.
- ***Humour.*** We know that humour is the best way to promote learning. 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 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 right through the rest of the book. (Check, for example, the exercises in Chapter 23. You'll see that we still require and demonstrate the plots and checks that were introduced in the early chapters.) This consistency helps reinforce these fundamental principles.

- ***The need to read.*** Students who plan just to skim the book may find our presentation frustrating. The important concepts, definitions, and sample solutions don't sit in little boxes. 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.

Mathematics

Mathematics can

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

Of these, we emphasize the first. 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 put off by equations, so we always provide a verbal description and a numerical example as well. Some theorems about Statistics are quite interesting, and many are important. Often, though, their proofs are not enlightening to introductory Statistics students and can distract the audience from the concepts we want them to understand. So we avoid them here.

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, they are often unnecessary. Today, virtually all statistics are calculated with technology, so there is little need for students to spend time summing squared deviations by hand. We have selected the equations that do appear for their focus on illuminating concepts and methods. Although these equations may be the best way to understand the concepts, they may not be optimal for hand calculation. When that happens, we give an alternative formula, better suited for hand calculation, for those who find following the process a better way to learn about the result.

Technology and Data

To experience the real world of Statistics, use modern technology to explore real data sets.

Technology. We assume that you are using some form of technology—a statistics package, a calculator, a spreadsheet, or some combination of these—in your Statistics course. We also assume that you'll put little emphasis on calculating answers by hand, even though we often show how. However, this is not a technology-heavy book. The role of technology in this book is to get the calculations out of the way so we can focus on statistical thinking. We discuss generic computer output, but we don't adopt any particular statistics software. We do offer guidance to help students get started on eight common software platforms: Excel®, Minitab®, JMP®, SPSS®, TI-83/84 Plus graphing calculators, StatCrunch®, and R®. *StatCrunch* can be accessed through *MyLab Statistics*, available from Pearson with the text. The **On the Computer** section at the end of most chapters is specific to the methods learned in that chapter.

Data. Because we use technology for computing, we don't limit ourselves to small, artificial data sets. You'll find some small data sets, but we also base examples and exercises on real data with a moderate number of cases—usually more than you would want to enter by hand into a program or calculator. Machine-readable versions of the data, appropriate for a variety of statistical packages, are included on the book's website, **www.pearsoncanada.ca/deveaux.**

New and Continuing Features

Enhancing Understanding

Learning Objectives and Connections. Each chapter starts with a brief summary of the chapter's key learning objectives, as well as a paragraph or two that point out the kinds of questions students will learn how to answer in the chapter and how this connects with or builds on earlier material in the text.

Each chapter ends with a **What Have We Learned?** summary, which includes new learning objectives and definitions of terms introduced in the chapter. Students can think of these as study guides.

In each chapter, our innovative **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.

Margin and in-text boxed notes. Throughout each chapter, boxed margin and in-text notes enhance and enrich the text. Boxes with essential or practical information are screened. Conversational notes that enhance the text and entertain the reader are unscreened.

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

By Hand. Even though we encourage the use of technology to calculate statistical quantities, we realize the pedagogical benefits of occasionally doing a calculation by hand. The By Hand boxes break apart the calculation of many simpler formulas to help the student through the calculation of a worked example.

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 often 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 as 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.

Learning by Example

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

Step-by-Step Examples: Think, Show, Tell. Step-by-Step examples repeat the mantra of Think, Show, and Tell in every chapter. These longer, worked examples guide students through the process of analyzing the problem with the general explanation on the left and the worked-out problem on the right. They emphasize the importance of thinking about a Statistics question (What do we know? What do we hope to learn? Are the assumptions and conditions satisfied?) and reporting our findings (the Tell step). The Show step contains the mechanics of calculating results and conveys our belief that it is only one part of the process.

The result is a better understanding of the concept, not just number crunching. In this edition, we've updated many Think, Show, Tell Step-by-Step examples with new demonstrations and data.

Testing Understanding

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.

Exercises. We've added section-specific single-concept exercises at the end of almost every section so students can be sure they have a clear understanding of each important topic before they're asked to tie them all together in more comprehensive exercises at chapter end. Chapter exercises have been updated with the most recent data available. Many come from news stories; some from recent research articles. Whenever possible, the data are available on MyLab Statistics and the Companion Website for the book so students can explore them further.

Technology

Data Sources. Most of the data used in examples and exercises are from real-world sources, and whenever we can, we include references to the Internet data sources used, often in the form of URLs. The data we use are usually on MyLab Statistics and the Companion Website. If you seek the data—or an updated version of the data—on the Internet, we try to direct you to a good starting point.

On the Computer. In the real world, Statistics is practiced with computers. We prefer not to choose a particular Statistics program. Instead, at the end of most chapters, we summarize what students can find in the most common packages, often with annotated output. We then offer specific guidance for several of the most common packages (Excel®, JMP®, Minitab®, R®, SPSS®, StatCrunch®, and TI-83/84 Plus[2]) to help students get started with the software of their choice.

[2]For brevity, we will write TI-83/84 Plus for the TI-83 Plus and/or TI-84 Plus. Keystrokes and output remain the same for the TI-83 Plus and the TI-84 Plus, so instructions and examples serve for both calculators.

Supplements

For Instructors

Instructor resources are password protected and available for download via the Pearson online catalogue at **http://catalogue.pearsoned.ca**.

Instructor's Solutions Manual. This resource provides complete, detailed, worked-out solutions for all the exercises in the textbook and is available through the online catalogue in both PDF and Word formats.

TestGen and Test Item File. For your convenience, our test bank is available in two formats. TestGen is a computerized testbank containing a broad variety of multiple-choice, short answer, and more complex problem questions. Questions can be searched and identified by question type, level of difficulty, and skill type (computational or conceptual). Each question has been checked for accuracy and is available in the latest version of TestGen software. This software package allows instructors to custom design, save, and generate classroom tests. The test program permits instructors to edit, add, or delete questions from the test bank; edit existing graphics and create new ones; analyze test results; and organize a database of tests and student results. This software allows for greater flexibility and ease of use. It provides many options for organizing and displaying tests, along with search and sort features. The same questions can also be found in a Test Item File available in Word format. The TestGen testbank and Test Item File can be downloaded from the online catalogue.

PowerPoint® Presentations. These PowerPoint® lecture slides provide an outline to use in a lecture setting, presenting definitions, key concepts, and figures from the textbook. These lecture slides can be downloaded from the online catalogue.

Image Library. The Image Library provides access to many of the images, figures, and tables in the textbook and is available to instructors on the online catalogue.

For Students

Student Solutions Manual. This solutions manual provides complete worked-out solutions to all of the odd-numbered exercises in the book, expanding on the answers provided in the Appendix at the back of the text.

Companion Website. The Companion Website provides additional resources (and data files) for instructors and students: **www.pearsoncanada.ca/deveaux.**

Technology Resources

MyLab™ Statistics Online Course (access code required)

- MyLab Statistics is a course management system that delivers **proven results** in helping individual students succeed.
- MyLab Statistics 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.
- MyLab Statistics' 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. You can determine which points of data you want to export, and then analyze the results to determine success.

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

- **Tutorial Exercises with Multimedia Learning Aids:** The homework and practice exercises in MyLab Statistics 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.
- **StatTalk Videos:** *24 Conceptual Videos to Help You Actually Understand Statistics*. Fun-loving statistician Andrew Vickers takes to the streets of Brooklyn, New York, to demonstrate important statistical concepts through interesting stories and real-life events. These fun and engaging videos will help students actually understand statistical concepts. Available with an instructor's user guide and assessment questions.
- **Getting Ready for Statistics:** A library of questions now appears within each MyLab Statistics to offer the developmental math topics students need for the course. These can be assigned as a prerequisite to other assignments, if desired.
- **Conceptual Question Library:** In addition to algorithmically regenerated questions that are aligned with your textbook, there is a library of 1000 Conceptual Questions available in the assessment manager that requires students to apply their statistical understanding.
- **StatCrunch®:** MyLab Statistics 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, MyLab Statistics includes access to **www.StatCrunch.com**, a website where users can access more than 15,000 shared data sets, conduct online surveys, perform complex analyses using the powerful statistical software, and generate compelling reports.

- **Statistical Software Support:** Knowing that students often use external statistical software, we make it easy to copy our data sets, both from the eText and the MyLab Statistics questions, into software such as StatCrunch, Minitab, Excel, and more. Students have access to a variety of support tools—Technology Tutorial Videos, Technology Study Cards, and Technology Manuals for select titles—to learn how to effectively use statistical software.

And, MyLab Statistics comes from a **trusted partner** with educational expertise and an eye on the future.

- Knowing that you are using a Pearson product means knowing that you are using quality content. That means that our eTexts are accurate and our assessment tools work. Whether you are just getting started with MyLab Statistics, or have a question along the way, we're here to help you learn about our technologies and how to incorporate them into your course.

To learn more about how MyLab Statistics combines proven learning applications with powerful assessment, visit **www.mystatlab.com** or contact your Pearson representative.

MathXL® for Statistics Online Course (access code required)

MathXL is the homework and assessment engine that runs MyLab Statistics. (MyLab Statistics is MathXL plus a learning management system.)

With MathXL for Statistics, 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 for Statistics, 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.
- Students can also access supplemental animations and video clips directly from selected exercises.

MathXL for Statistics is available to qualified adopters. For more information, visit our website at **www.mathxl.com**, 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 more than 15,000 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 MyLab Statistics, and StatCrunch is available by itself to qualified adopters. StatCrunch Mobile is now available to access from your mobile device. For more information, visit our website at **www.StatCrunch.com**, or contact your Pearson representative.

Acknowledgments

A very big thanks goes to Shivon Sue-Chee for her hard work on the exercise sections, data sets, and appendix solutions. Her hard work and conscientiousness were, as always, a marvel. Also, a big thanks to Aaron Springford and Andy Leung for their very precise work in updating the Instructor's Solutions Manual.

Thank you also to the team at Pearson who supported the revision of this text throughout the writing and production process including Paul Donnelly, Senior Content Developer; Pippa Kennard, Project Manager; Emily Dill, Content Manager; and Mark Grzeskowiak, Acquisitions Editor.

Finally, thank you to the reviewers who provided feedback on our work and helped guide our plans for this new edition, including:

James Adcock, *University of Waterloo*
Nancy Chibry, *University of Calgary*
Torben Drewes, *Trent University*
Tracy Finlay, *Fleming College*
Renzo Gomez, *Durham College*
Erin Kirk, *Fleming College*
Michele Millar, *Mount Saint Vincent University*
Julie Orsini, *Durham College*
Scott Robison, *University of Calgary*
Tim Swartz, *Simon Fraser University*
Asokan Mulayath Variyath, *Memorial University of Newfoundland*

This page intentionally left blank

Index of Application

Note: (E) = Exercise, (FE) = For Example, (JC) = Just Checking, (IE) = in-text example (SBS) = Step-by-Step

Agriculture

Arts and Literature

Biology

Business

Economics

Education

Entertainment

Environment

Finance

Food and Drinks

Games

Genetics

Health

Law

Media

Miscellaneous

Music

People

Politics

Psychology

Science

Social Issues

Sports

Surveys

Technology

Telecommunication

Transport

Web

STATS

DATA AND MODELS

This page intentionally left blank

CHAPTER

1 Stats Starts Here

LEARNING OBJECTIVES

1. Recognize the vital role of Statistics in answering real-world questions involving variation and uncertainty.
2. Understand that data are values with their context.
3. Identify cases and variables.
4. Identify whether a variable is being used as categorical or quantitative.

Connections

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 can be fun, though you might not guess it from some of the book covers popping up now and then at Indigo, with titles like *Statistics for the Terrified* and *Statistics without Tears*. Well, we certainly hope to calm your fears and hold back the tears while helping you learn to think clearly when dealing with data.

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

"But where shall I begin?" asked the White Rabbit.
"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 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.[1]

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 them into a powerful, effective approach to understanding the world through data.

Data vary. Ask different people the same question and you'll get a variety of answers. Even when you ask the same person the same question at different times, you may have a variety of answers! Statistics helps us to make sense of the world 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*,[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

[1] *Statistical Science* would be a more appropriate name. Statistics is indeed a science and integral to the scientific method. Its methodology extends far beyond the stereotype of data compilation and tabulation, and calculation of summary *statistics*. It overlaps with a rather new field of knowledge called *Data Science.*

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

Q: What is Statistics?
A: Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world.
Q: What are statistics?
A: Statistics (plural) are particular calculations made from data.
Q: So what is data?
A: You mean, "What are data?" Data is the plural form. The singular is datum.
Q: OK, OK, so what are data?
A: Data are values along with their context.

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

Statistics is about *variation*.
Data vary because we don't see everything, and because even what we do see and measure, we measure imperfectly.
But Statistics helps us understand and model this variation so that we can see through it to the underlying truths and patterns.[4]

something both beautiful and valuable. What we're going to talk about in this book is how you can mine your own data and gain valuable insights about the world.

- In December 1992, British engineer Neil Papworth (now a Montrealer) texted "Happy Christmas' to a company director. That was the first text message ever sent, followed by trillions more in ensuing years, as the number of text messages and messaging apps have grown at a ridiculous pace. Some of these messages were sent or read while the sender or the receiver was 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.[3] The results were striking. The texting drivers actually responded more slowly and were more dangerous than those who were above the legal limit for alcohol. In this book, you'll learn to design and analyze experiments like this.

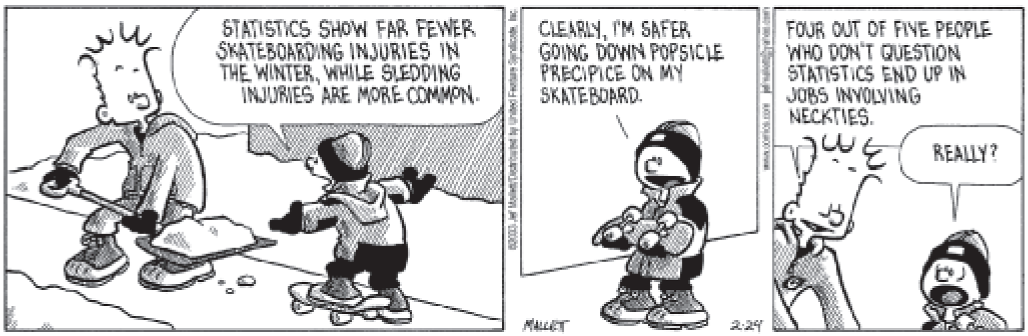
How do we assess the risk of genetically engineered foods being considered by the Canadian Food Inspection Agency or the safety and effectiveness of new drugs submitted to Health Canada for approval? How can we predict the number of new cases of AIDS by regions of the country or the number of customers likely to respond to a sale at the mall? Or determine whether enriched early education affects later performance of school children, and whether vitamin C really prevents illness? Statistics is all about answering real-world questions such as these.

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.

Statistics in a Word

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

Economics is about … *Money (and why it is good).*
Psychology: *Why we think what we think (we think).*
Paleontology: *Previous Life*.
Biology: *Life.*
Religion: *After Life*
Anthropology: *Who?*
History: *What, where, and when?*
Philosophy: *Why?*
Engineering: *How?*
Accounting: *How much?*
In such a caricature, Statistics is about … *Variation.*



FRAZZ © 2003 Jef Mallett. Distributed by Universal Uclick. Reprinted with permission. All rights reserved.

[3] "Text Messaging During Simulated Driving," Drews, F. A. et al., Human Factors: hfs.sagepub.com/content/51/5/762
[4] Seeing the forest *through* the trees

Section 1.1 Exercises

1. **Grocery shopping** Many grocery store chains offer customers a card they can scan when they check out and offer discounts to people who do so. To get the card, customers must give information, including a mailing address and e-mail address. The actual purpose is not to reward loyal customers but to gather data. What data do these cards allow stores to gather, and why would they want that data?

2. **Online shopping** Online retailers such as Amazon.com keep data on products that customers buy, and even products they look at. What does Amazon hope to gain from such information?

1.2 Data

©Dorling Kindersley

Amazon.com opened for business in July 1995, billing itself as "Earth's Biggest Bookstore." By 1997, Amazon had a catalogue of more than 2.5 million book titles and had sold books to more than 1.5 million customers in 150 countries. In 2015, the company's sales reached $107.0 billion (a 20% increase from the previous year and more than double of 2011 sales). Amazon has sold a wide variety of merchandise, including a $400,000 necklace, yak cheese from Tibet, and the largest book in the world (see photo at left). Amazon.ca launched in June 2002, following up in April 2011 with a new shipping facility built in Mississauga, Ontario. How did Amazon become so successful and how can it keep track of so many customers and such a wide variety of products? The answer to both questions is *data.*

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

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

105-2686834-3759466	Quebec	Nashville	Kansas	10.99	819	N	B00000I5Y6	Katherine H.
105-9318443-4200264	Nova Scotia	Orange County	Boston	16.99	902	Y	B000002BK9	Samuel P.
105-1872500-0198646	Alberta	Bad Blood	Chicago	15.98	403	N	B000068ZVQ	Chris G.
103-2628345-9238664	Ontario	Let Go	Mammals	11.99	416	N	B000001OAA	Monique D.
002-1663369-6638649	Quebec	Best of Kansas	Kansas	10.99	819	N	B002MXA7Q0	Katherine H.

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

Order Number	Name	Province	Price	Area Code	Previous Album Download	Gift?	ASIN	New Purchase Artist
105-2686834-3759466	Katherine H.	Quebec	10.99	819	Nashville	N	B0000015Y6	Kansas
105-9318443-4200264	Samuel P.	Nova Scotia	16.99	902	Orange County	Y	B000002BK9	Boston
105-1372500-0198646	Chris G.	Alberta	15.98	403	Bad Blood	N	B000068ZVQ	Chicago
103-2628345-9238664	Monique D.	Ontario	11.99	416	Let Go	N	B0000010AA	Mammals
002-1663369-6638649	Katherine H.	Quebec	10.99	819	Best of Kansas	N	B002MXA7Q0	Kansas

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

What information would provide a **context**? Newspaper journalists know that the lead paragraph of a good story should establish the "Five W's": *who, what, when, where,* and (if possible) *why*. Often, we add *how* to the list as well. The answers to the first two questions are essential. If we don't know *what* attributes or characteristics are measured and *who* they are measured on, the values displayed 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. Those characteristics (called variables—see Section 1.2) constitute the "*what*" of the study. The *"who"* or the cases go by different names, depending on the situation.

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

Data trumps intuition
Amazon monitors and evolves 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 … Data is king at Amazon."

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, both in the first and the last row. Each row of data is information about a single case, and that case is often identified in the leftmost column. 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 may also 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 that we recommend you adopt.

The first step of any data analysis is to know what you are trying to accomplish and what you want to know (*why?*). 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.

Wavebreakmedia/Shutterstock

For Example IDENTIFYING THE "WHO"

In December 2013, *Consumer Reports* published an evaluation of 46 tablets from a variety of manufacturers (http://www.consumerreports.org/cro/magazine/2013/12/best-electronics-brands/index.htm).

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 46 tablets, which are the "Who" for these data. Each tablet selected represents all similar tablets offered by that manufacturer.

Section 1.2 Exercises

3. Super Bowl Sports announcers love to quote statistics. During the Super Bowl, they particularly love to announce when a record has been broken. They might have a list of all Super Bowl games, along with the scores of each team, total scores for the two teams, margin of victory, passing yards for the quarterbacks, and many more bits of information. Identify the *who* in this list.

4. Nobel laureates The web page at www.nobelprize.org/nobel_prizes/peace/laureates/ allows you to look up all the Nobel prizes awarded in any year. The data are not listed in a table. Rather you drag a slider to a list of prize categories and then see a list of the awardees for each category for every year. Describe the *who* in this scenario.

1.3 Variables

The characteristics (the *"what"*) 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, Province, 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 *Province*. 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.)

Donatas1205/Fotolia

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 dialled. New York City was assigned 212, Chicago 312, and Los Angeles 213, while Halifax was 902, Quebec City 418, and Vancouver 604. 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 now become just categories.

Descriptive responses to questions are often categories. For example, the responses to the questions "Who is your cell phone provider?" or "What is your marital status?" yield categorical values. 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 overnight air. *Counting is a natural way to summarize and learn about a categorical variable* like *Shipping Method*. The counts though are just summaries of the actual raw categorical data, which looks like a long list of shipping methods, one shipping method for each order. Chapter 2 discusses summaries and displays of categorical variables more fully.

"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"**

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, kilometres 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, Canadian dollars, U.S. dollars, or Chinese yuan.

Aside from the process of measurement, quantitative variables may also arise from the process of counting. Counts also measure the amounts of things. How many songs are on your iPod? How many siblings do you have? To measure these quantities, we would naturally count, instead of measure, and we'd consider the units to be "number of …" or just "counts." Amazon might be interested in the number of teenage customers visiting its site each month, in order to track customer growth and forecast sales. So, Who = months and What = # teens (whereas when Amazon counts the number of ground, second-day air, and overnight shipments during some time period, these counts are just summaries for the categorical variable *Shipping Method*, where Who = purchases and What = method). Chapter 3 discusses summaries and displays for quantitative variables.

Sometimes a variable with numeric values may be treated as categorical. Amazon could record your *Age* in years. That seems quantitative, and it would be if the company wanted to know the average age of those customers who visit their site after 3 a.m. But suppose Amazon wants to decide which album to feature on its site when you visit. Then placing your age into one of the categories *Child*, *Teen*, *Adult*, or *Senior* might be more useful. You can always convert a quantitative variable to a categorical one, by simply breaking up the range of values into several intervals. There is some loss of information, but if this facilitates a particular method of analysis, it may still be worthwhile doing.

Privacy and the Internet You have many identifiers: a social insurance number, a student ID number, 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 behaviour 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.

Identifiers

For a categorical variable like *Sex*, each individual is assigned one of two possible values, say *M* or *F*. But for a variable with ID numbers, such as *Student ID*, each individual receives a unique value. A variable like this, used solely to uniquely identify each of the cases, is called 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. However, they 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. Many large databases are *relational* databases. In a relational database, different data tables link to one another by matching identifiers. In the Amazon example, *Transaction Number* is clearly an identifier. What about a CD's ASIN? It depends on the purpose of the study. If we want to determine the best-selling CD, then its ASIN is a categorical variable of interest and we would count how often each ASIN appears in a row or transaction. But if we are examining the range of prices for CDs, the ASIN could be used as an identifier variable.

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 quantitative. To treat it as quantitative, the teacher will have to imagine that it has "educational value units" or some similar arbitrary construct. Because there are no natural units, any conclusions she draws should be viewed with considerable caution! Variables that report order without natural units are 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. Ordinal variables are often viewed as a subclass of categorical variables, along with nominal (no natural ordering) variables, since nominal and ordinal variables both arise through a process of categorizing your cases.

For Example IDENTIFYING "WHAT" AND "WHY" OF TABLETS

RECAP: A *Consumer Reports* article about 46 tablets lists each tablet's manufacturer, price, battery life (hrs.), operating system (iOS/Android/Windows), and overall performance 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)
- performance 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 seven-time winner and once again set a new record for the fastest average speed. And on January 17, 2013, on the Oprah Winfrey show, he made history yet again, confessing to a long history of extensive doping! You can find data on all the Tour de France races in the data set **Tour de France 2015**. Here are the first three and last eight lines of our data set. Keep in mind that the entire data set has over 100 entries. And as promised in our second edition, the Armstrong entries, like his medals, have vanished with our third edition!

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
⋮										
2007	Alberto Contador	Spain	24	Discovery Channel	91.00.26	38.97	20	3547	189	141
2008	Carlos Sastre	Spain	33	CSC-Saxo Bank	87.52.52	40.50	21	3559	199	145
2009	Alberto Contador	Spain	26	Astana	85.48.35	40.32	21	3460	180	156
2010	Andy Schleck	Luxembourg	25	Saxo Bank	91.59.27	39.590	20	3642	180	170
2011	Cadel Evans	Australia	34	BMC	86.12.22	39.788	21	3430	198	167
2012	Bradley Wiggins	Great Britain	32	Sky	87.34.47	39.827	20	3488	198	153
2013	Christopher Froome	Great Britain	28	Sky	83.56.40	40.551	21	3404	198	169
2014	Vincenzo Nibali	Italy	29	Astana	89.56.06	40.735	21	3663.5	198	164
2015	Christopher Froome	Great Britain	30	Sky	84.46.14	39.64	21	3360.3	198	160

Section 1.3 Exercises

5. **Grade levels** A person's grade in school is generally identified by a number.
   a) Give an example of a *why* in which grade level is treated as categorical.
   b) Give an example of a *why* in which grade level is treated as quantitative.

6. **ZIP codes** Unlike in Canada, the U.S. Postal Service uses five-digit ZIP codes to identify locations to assist in delivering mail.
   a) In what sense are ZIP codes categorical?
   b) Is there any ordinal sense to ZIP codes? In other words, does a larger ZIP code tell you anything about a location compared to a smaller ZIP code?

7. **Affairs** A May 2015 Gallup Poll Canada question asked, "Please tell me whether you personally believe that in general it is morally acceptable, or morally wrong for married men and women having an affair?" The possible responses were "Morally acceptable," "Morally wrong," "Depends on situation," "Not a moral issue," and "No opinion." What kind of variable is the response?

8. **Economic outlook** An April 2016 Gallup Poll Canada asked Canadians, "What is your opinion about the local economy six months from now?" The possible responses were "Grows strong," "Remains the same," "Gets weaker," and "Don't know." What kind of variable is the response?

9. **Medicine** A pharmaceutical company conducts an experiment in which a subject takes 100 mg of a substance orally. The researchers measure how many minutes it takes for half of the substance to exit the bloodstream. What kind of variable is the company studying?

10. **Stress** A medical researcher measures the increase in heart rate of patients under a stress test. What kind of variable is the researcher studying?

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 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 favourite statistics or spreadsheet program to reformat the data into variables. You will also probably want to remove commas from large numbers and extra symbols such as money indicators ($, ¥, £); few statistics packages can handle these.

WHAT CAN GO WRONG?

- **Don't label a variable as categorical or quantitative without thinking about the 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 colours 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.

  Be very cautious (if not skeptical) when an ordinal variable is converted to a numerical scale for analysis.

What Have We Learned?

Understand that data are information (recorded as numbers or labels) in a 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.

We treat variables in two basic ways, as *categorical* (sometimes called qualitative) or *quantitative*.

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