BUSINESS STATISTICS

SHARPE | DE VEAUX | VELLEMAN | WRIGHT

FOURTH CANADIAN EDITION

To my loving family for their patience and support
—Norean

To my father, whose daily stories informed me how the world of business really worked, and to my family, for giving me the love and support that made this book possible
—Dick

To my father, who taught me about ethical business practice by his constant example as a small businessman and parent
—Paul

To Mina, Ramin, Leila, Norman, Rebecca, and Allison
—David

Meet the Authors

Norean Radke Sharpe (Ph.D. University of Virginia) is Dean and the Joseph H. and Maria C. Schwartz Distinguished Chair at the Peter J. Tobin College of Business at St. John's University. As the chief academic officer of the Tobin College of Business, she is responsible for the curriculum for 2500 undergraduate business majors and 600 graduate students in one of seven M.S./M.B.A. programs, all supported by more than 150 faculty and staff on the Manhattan, Queens, Staten Island, and Rome, Italy, campuses. Within the Tobin College is the Center for Enterprise Risk Management, the Applied Finance Institute, and the Global Business Stewardship Center, as well as the acclaimed School of Risk Management, Insurance, and Actuarial Science.

Dr. Sharpe is an accomplished scholar, with 30 years of teaching experience at Yale University, Bowdoin College, Babson College, and Georgetown University—and with more than 30 scholarly publications in analytics and statistics education. Her research interests include time series analysis, forecasting, analytics, and women's roles in entrepreneurship in the Middle East. Dr. Sharpe earned her B.A. from Mt. Holyoke College, M.S. from the University of North Carolina, and Ph.D. in Systems Engineering from the University of Virginia.

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

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

Richard De Veaux and Paul Velleman have authored successful books in the introductory college and AP High School market with David Bock, including *Intro Stats*, Fifth Edition (Pearson, 2018); *Stats: Modeling the World*, Fifth Edition (Pearson, 2019); and *Stats: Data and Models*, Fourth Edition (Pearson, 2016).

David Wright combines an Engineering Ph.D. from Cambridge University, UK, with his current position as Full Professor at the University of Ottawa's Telfer School of Management to provide a business perspective on renewable energy. Dr. Wright has taught in universities in North America, Europe, and Africa and has experience in government and in industry. In government, he has developed statistical models to evaluate the impact of industrial society on natural resource depletion. In industry, he has worked with solar power developers on the economic viability and risk assessment of solar power projects. He also has experience in the telecommunications industry on the impact of new technology on business. His university teaching and research includes the economics of solar energy, business statistics, and the smart electricity grid. In his research, he collaborates with professors in engineering and with industrial partners. He is cited in *Who's Who in the World, Who's Who in Canadian Business,* and *Who's Who in Science and Engineering.*

Brief Contents

Contents

Preface

The question that motivates a business student's study of Statistics is "How can I make better decisions?" As entrepreneurs and consultants, we know that in today's data-rich environment, knowledge of Statistics is essential to survive and thrive in the business world. But, as educators, we've seen a disconnect between the way Business Statistics is traditionally taught and the way it should be used in making business decisions. In *Business Statistics*, we try to narrow the gap between theory and practice by presenting relevant statistical methods that will empower business students to make effective, data-informed decisions.

Of course, students should come away from their Statistics course knowing how to think statistically and how to apply Statistics methods with modern technology. But they must also be able to communicate their analyses effectively to others. When asked about Statistics education, a group of CEOs from *Fortune* 500 companies recently said that although they were satisfied with the technical competence of students who had studied Statistics, they found the students' ability to communicate their findings to be woefully inadequate.

Our "Plan, Do, Report" rubric provides a structure for solving business problems that mimics the correct application of statistics to solving real business problems. Unlike many other books, we emphasize the often neglected thinking (Plan) and communication (Report) steps in problem solving in addition to the methodology (Do). This approach requires up-to-date, real-world examples and data. So we constantly strive to illustrate our lessons with current business issues and examples.

We've been delighted with the reaction to previous editions of *Business Statistics*. We continue to update examples and exercises so that the story we tell is always tied to the ways Statistics informs modern business practice. We teach with real data whenever possible, so we've updated data in the Section and Chapter Exercises. New examples reflect current stories in the news and recent economic and business events.

Statistical Thinking

Our mission for every edition has been to write a modern Business Statistics text that addresses the importance of *statistical thinking* when making business decisions and that acknowledges how Statistics is actually used in business.

Statistics is practised with technology. This insight informs everything, from our choice of forms for equations (favouring intuitive forms over calculation forms) to our extensive use of real data. But most important, understanding the value of technology allows us to focus on teaching statistical thinking rather than just calculation. The questions that motivate each of our hundreds of examples are geared not to the question "How do you find the answer?" but rather to the question "Does your data satisfy the necessary assumptions and how do you apply the result to a business situation?" This focus on statistical thinking ties the chapters of the book together.

Our Goal: Read This Book!

The best textbook in the world is of little value if it isn't read. Here are some of the ways we made *Business Statistics* more approachable:

- ***Readability***. We strive for a conversational, approachable style, and we introduce anecdotes to maintain interest. While using the First Canadian Edition, instructors reported (to their amazement) that their students read ahead of their assignments voluntarily. Students write to tell us (to *their* amazement) that they actually enjoy the book.

- ***Focus on assumptions and conditions.*** More than any other textbook, *Business Statistics* emphasizes the need to verify assumptions when using statistical procedures. We emphasize this focus throughout the examples and exercises. We make every effort to provide templates that reinforce the practice of checking assumptions and conditions, rather than simply rushing through the computations. Business decisions have consequences. Blind calculations open the door to errors that could easily be avoided by taking the time to graph the data, check assumptions and conditions, and then check again that the results make sense.
- ***Emphasis on graphing and exploring data.*** Our consistent emphasis on the importance of displaying data is evident from the first chapters devoted to understanding data to the sophisticated model-building chapters at the end of the book. Examples often illustrate the value of examining data graphically, and the exercises reinforce this concept. Graphics reveal structures, patterns, and occasional anomalies that could otherwise go unnoticed. The sight of patterns displayed graphically often raises new questions and informs both the path of a resulting statistical analysis and the ensuing business decisions. The graphics that appear throughout the book also demonstrate that the simple structures that underlie even the most sophisticated statistical inferences are the same ones we look for in the simplest examples. That helps to tie the concepts of the book together to tell a coherent story.
- ***Consistency.*** Having taught the importance of plotting data and checking assumptions and conditions, we are careful to model that behaviour throughout the book. (Check the exercises in the chapters on multiple regression or time series and you'll find us still requiring and demonstrating the plots and checks that were introduced in the early chapters.) This consistency helps reinforce these fundamental principles and provides a familiar foundation for the more sophisticated topics.

Coverage

We were guided in our choice of topics by the GAISE 2016 (Guidelines for Assessment and Instruction in Statistics Education) Report, which emerged from extensive studies exploring how students best learn Statistics (**https://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf**). Those recommendations have been officially adopted and recommended by the American Statistical Association and urge (among other detailed suggestions) that Statistics education should achieve the following goals:

1. Teach statistical 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; and
6. Use assessments to improve and evaluate student learning.

With respect to the order of topics, we followed the principle that a coherent introductory course should be designed so that concepts and methods fit together in a stepwise progression to provide a new understanding of how reasoning with data can uncover new and important truths. For example, we teach inference concepts with proportions *first* and then with means. Most students have had exposure to proportions through polls and advertising. And by starting with proportions, we can teach inference with the Normal model and then

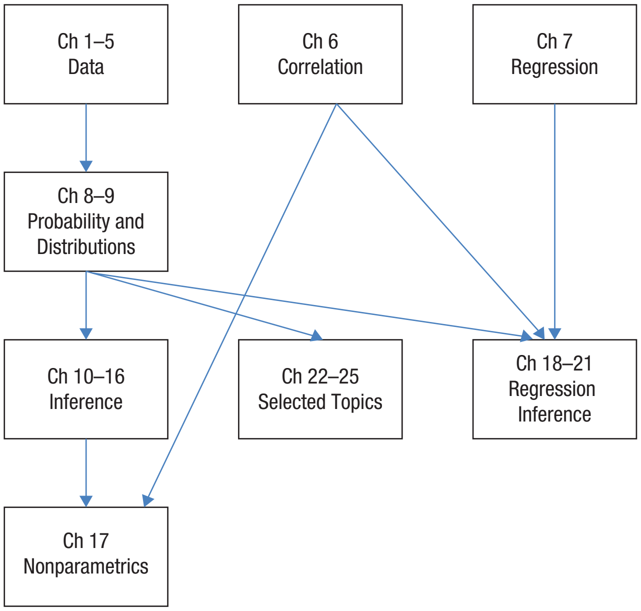

Figure 1 Visual map of the links between chapters

introduce inference for means with the Student's *t* distribution. We introduce the concepts of association, correlation, and regression *early* in *Business Statistics*. Our experience in the classroom shows that introducing these fundamental ideas early makes Statistics useful and relevant, even at the beginning of the course. Later in the semester, when we explore data through inference, it feels natural and logical to build on the fundamental concepts learned earlier.

Syllabus Flexibility

Many instructors prefer to teach topics in a different sequence than the one presented in the textbook. In order to assist you with your decision, Figure 1 is a diagram that illustrates the dependency among chapters.

The subject of Business Statistics is sometimes taught in a single semester and other times taught over the course of two semesters. Table 1 offers one suggestion for the way in which chapters can be divided between two semesters.

	Core Topics			Optional Topics			
	Data	**Regression**	**Probability Distributions**	**Inference**	**Nonparametrics**	**Multiple Regression**	**Selected Topics**
First Semester	Ch 1–5	Ch 6–7	Ch 8–9				Ch 22–25
Second Semester		Ch 18–19		Ch 10–16	Ch 17	Ch 20–21	

Table 1 Chapter selection

Features

A textbook isn't just words on a page—instead, it's the cumulation of many features that form a big picture. The features in *Business Statistics* are designed to provide a real-world context for concepts, to help students to apply these concepts, to promote problem solving, and to integrate technology—all in the name of helping students to more readily identify the key themes the book is trying to teach.

Motivating Vignettes. Each chapter opens with a motivating vignette, often taken from the authors' consulting experiences. These descriptions of companies—such as Bell Canada, Sport Chek, Rogers, Intact Financial Corp., Ipsos Reid, PotashCorp of Saskatchewan, Canada's Wonderland, and Loblaw—enhance and illustrate the story of each chapter and show how and why statistical thinking is so vital to modern business decision making. We analyze data from or about the companies in the motivating vignettes throughout the chapter.

FOR EXAMPLE

For Examples. Nearly every section of every chapter includes a focused example that illustrates and applies the concepts or methods of that section. The best way to understand and remember a new theoretical concept or method is to see it applied in a real-world business context right away. That's what these examples do throughout the book.

PLAN

DO

REPORT

Step-by-Step Guided Examples. The answer to a statistical question is almost never just a number. Statistics is about understanding the world and making better decisions with data. To that end, some examples in each chapter are presented as Guided Examples. A thorough solution is modelled in the right column while commentary appears in the left column. The overall analysis follows our innovative **Plan, Do, Report** template. That template begins each analysis with a clear question about a business decision and an examination of the data available (**Plan**). It then moves to calculating the selected statistics (**Do**). Finally, it concludes with a **Report** that specifically addresses the question. To emphasize that our goal is to address the motivating question, we present the **Report** step as a business memo that summarizes the results in the context of the example and states a recommendation if the data are able to support one. To preserve the realism of the example, whenever it is appropriate we include limitations of the analysis or models in the concluding memo, as would be required when writing a report for management.

WHAT CAN GO WRONG?

What Can Go Wrong? Each chapter contains an innovative section called "What Can Go Wrong?" which highlights the most common statistical errors and the misconceptions about Statistics. The most common mistakes for the new user of Statistics involve misusing a method—*not* miscalculating a statistic. Most of the mistakes we discuss have been experienced by the authors in a business context or in a classroom situation. 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. In this spirit, some of our exercises probe how, and why, common errors tend to arise.

NOTATION ALERT

Notation Alert. Throughout this book, we emphasize the importance of clear communication. Proper notation is part of the vocabulary of Statistics, but it can be daunting. We all know that in Algebra, n can stand for any variable, so it may be surprising to learn that in Statistics, n is reserved for the sample size. Statisticians dedicate many letters and symbols for specific meanings (e.g., the letters b, e, n, p, q, r, s, t, and z, along with many Greek letters, all carry special connotations). Our "Notation Alerts" clarify which letters and symbols statisticians use and the purpose of each letter and symbol.

JUST CHECKING

Just Checking. It is easy to start nodding in agreement without really understanding, so we ask questions at points throughout the chapter. These questions are designed to conduct a quick check of whether or not students have properly understood a section; most involve very little calculation, and the answers are given in Appendix A. The questions can also be used to motivate class discussion.

Optional Math Box

Optional Math Boxes. In many chapters we present the mathematical underpinnings of the statistical methods and concepts. We set proofs, derivations, and justifications apart from the narrative in "Optional Math Boxes," so the underlying mathematics is available for those who want greater depth, but the text itself presents the logical development of the topic at hand using a minimal amount of mathematics.

ETHICS IN ACTION

Ethics in Action. Statistics involves more than simply plugging numbers into formulas; most statistical analyses require a fair amount of judgment. When faced with these sorts of important judgments, the best advice we can offer is to make an honest and ethical attempt to address the appropriate business issue. The chapter-specific *Ethics in Action* boxes illustrate some of the judgments needed when conducting statistical analyses, identify possible errors, link the issues to the American Statistical Association's Ethical Guidelines, and then propose ethically and statistically sound alternative approaches.

WHAT HAVE WE LEARNED?

Learning Objectives and What Have We Learned? Each chapter begins with a specific list of learning objectives and ends by relating the objectives to the chapter summary (i.e., the "What Have We Learned?" section). We review the concepts, define the terms introduced in the chapter, and list the skills that form the core message of the chapter. The "What Have We Learned?" sections make excellent study guides: the student who understands the concepts in the summary, knows the terms, and practises the skills correctly is better prepared to apply statistics to the world of business.

Technology Help

Technology Help. At the end of each chapter, we summarize what students can find in the most common software, often with annotated output. We then offer specific guidance for Excel, Minitab, SPSS, and JMP, formatted in easy-to-read sections. This advice is intended not to replace the documentation that accompanies the software, but rather to point the way and provide startup assistance.

MINI case studies

Mini Case Studies. Each chapter includes Mini Case Studies that ask students to conduct an analysis based on a real business situation. Students define the objective, plan the process, complete the analysis, and report a conclusion. An ideal way for students to write up their work is the "Plan/Do/Report" format described above and used in each chapter. Data for the Mini Case Studies are available on the MyLab Statistics site and are formatted for use with various technologies.

Case Studies. Parts 1, 2, and 3 of the book have a Comprehensive Case Study on MyLab Statistics. Students are given realistically large data sets (also on the MyLab Statistics site) and challenged to respond to open-ended business questions using the data. Students have the opportunity to bring together methods they have learned in the chapters included in that part (and indeed, throughout the book) to address the issues raised. Students will be required to use a computer to manipulate the large data sets that accompany these Case Studies.

EXERCISES

Section Exercises. The Exercises for each chapter begin with a series of straightforward exercises targeted at the topics in each chapter section. This is the place to check understanding of specific topics. Because the exercises are labelled by section, turning back to the right part of the chapter to clarify a concept or review a method is easy.

Chapter Exercises. These exercises are designed to be more realistic than the Section Exercises and to lead to conclusions about practical management situations. The Chapter Exercises may combine concepts and methods from different sections. We've worked hard to make sure that they contain relevant, modern, and realistic business situations. Whenever possible, the data are on the MyLab Statistics site (always in a variety of formats) so they can be explored further. Often, we pair the exercises so that each odd-numbered exercise (with answers that appear at the end of the book) is followed by an even-numbered exercise on the same Statistics topic.

Ⓣ The exercises marked with a data set icon in the margin indicate that the data are provided on the MyLab Statistics site.

Data and Sources. Most of the data used in examples and exercises stem from real-world sources. Whenever possible, we present the original data as we collected it. Sometimes, due to concerns about confidentiality or privacy, we had to change the values of the data or the names of the variables slightly, always being careful to keep the context as realistic and true to life as possible. Whenever we can, we include references to internet data sources. As internet users know well, URLs often break as websites evolve. To minimize the impact of such changes, we point as high in the address tree as is practical, so it may be necessary to search down into a site to find the data. Moreover, the data online may change as more recent values become available. The data we use are usually posted on the MyLab Statistics site.

Acknowledgements

Many thanks to the reviewers of this Fourth Canadian Edition, in alphabetical order:

Dr. Philipp Afèche, University of Toronto
Dr. Behrouz Bakhtiari, McMaster University
David Dobson, University of the Fraser Valley
Amanda Duncan, Georgian College
Martha Essak, University of British Columbia
Bruno Fullone, George Brown College
Gina Kerr, St. Lawrence College
Dr. Mohammad Mahbobi, Thompson Rivers University
Bob Saarimaki, Georgian College
Sofia Silberberg, Algoma University

Part 1 EXPLORING AND COLLECTING DATA

123RF

1

An Introduction to Statistics

LEARNING OBJECTIVES

In this chapter we show you how statistics is useful in business and why it will be increasingly in demand in the 21st century. After reading and studying this chapter, you should be able to:

1. Identify the importance of understanding statistics

The graphs and tables shown here are the daily bread and butter of investment managers and stock brokers. They're full of "statistics." Obviously this kind of information is important to them, but is this what Statistics is all about? Well, yes and no. This page may contain a lot of facts, but as we'll see, Statistics is much more interesting and rich than building and assessing graphs and tables.

Most companies have large databases, but there's not much point in having all that information sitting there unless we can analyze it. In the 20th century, we figured out how to store information and index it so that we can retrieve the items we want. The focus in the 21st century is on analyzing this information and using it to make effective business decisions. The field of "data analytics" is worth hundreds of billions of dollars, and it's growing at about 10% per year;[1] much of that analysis is statistical.

As a manager, the decisions you make based on data will chart the future course of your organization. You'll want to be able to interpret the data that surrounds you and come to your own conclusions. And you'll find that studying Statistics is much more important and enjoyable than you thought.

[1] Special report: Managing information: Data, data everywhere. (2010, February 25). *The Economist*.

LO 1 1.1 So What Is Statistics?

It is the mark of a truly intelligent person to be moved by statistics.

—George Bernard Shaw

Statistics is the basis for the global economy of the 21st century. If you didn't expect that answer, or if it sounds a bit grandiose, consider this: The global economy has undergone several dramatic changes over the years, as illustrated in Figure 1.1.

1. *The agricultural revolution*. We produced more food by farming than by hunting and gathering.
2. *The 19th-century industrial revolution*. Factories and mass production gave us a vast array of consumer and industrial products.
3. *The 20th-century information revolution*. Technology gave us a diverse range of electronic products, made our industry more efficient, and greatly increased the amount of information at our disposal.

But how can we make sense of all the data produced by the information revolution? Enter the next stage.

4. *The 21st-century data analytics revolution*. With vast volumes of information on hand, the challenge for the 21st century is extracting meaning from it all—and a key way of doing so is through statistical analysis.

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 quantities calculated from data.

Q: So what is data?

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

Q: So what are data?

A: Data are values, along with their context.

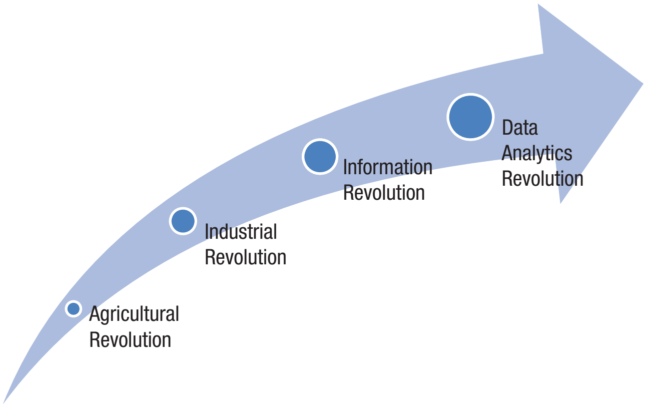

Figure 1.1 Revolutions in business.

Data analytics refers to the statistical analysis of large amounts of data in order to sift out the key information needed for corporate planning. Data analytics is becoming so powerful that some commentators claim it polarizes the labour market into "lousy and lovely jobs." And as *The Globe and Mail* put it, "The lovely jobs are why we should all enroll our children in statistics courses."[2]

Let's now look at some examples of what statistics can do for us. Most 20th-century applications of statistics continue to be important today, and some applications are new with the data analytics revolution of this century. So we'll start with the applications common to the 20th and 21st centuries, move on to what's new in this century, and then describe the cutting-edge applications that continue to be a challenge. As you read these examples, you can put them in context using Figure 1.2.

20th- and 21st-Century Statistics

Analyzing Large Amounts of Data

We've always used statistics to analyze both large and small amounts of data. We analyze large databases—for example, stock market and interest-rate data—for patterns that can identify what factors are associated with, say, an increase in share

[2] From The Globe and Mail by Chrystia Freeland. Published by The Globe and Mail Inc, © 2012.

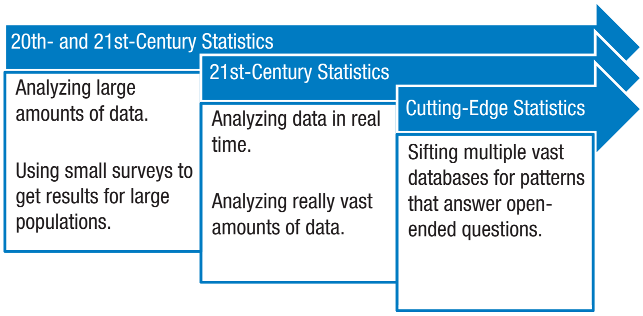
20th- and 21st-Century Statistics
Analyzing large amounts of data.
Using small surveys to get results for large populations.
21st-Century Statistics
Analyzing data in real time.
Analyzing really vast amounts of data.
Cutting-Edge Statistics
Sifting multiple vast databases for patterns that answer open-ended questions.

Figure 1.2 Trends in the use of statistical analysis.

prices or a lowering of interest rates. Similarly, retail firms like Loblaw and Future Shop analyze trends in retail sales, and insurance companies analyze trends in claims. We hope this text will empower *you* to draw conclusions from data and to make valid business decisions in response to such questions as

- Do aggressive, "high-growth" mutual funds really have higher returns than more conservative funds?
- Do your customers have common characteristics, and do they choose your products for similar reasons? And more importantly, are those characteristics the same among people who *aren't* your customers?
- What is the effect of advertising on sales?

Analyzing Small Amounts of Data

Drawing conclusions from small amounts of data is important, too. Indeed, one of the powers of statistical analysis is its ability to survey a small sample and generalize the results to a much larger population. (We talk more about sampling in Chapters 3 and 10, and the movement from the specific to the general is a theme we revisit throughout this book.) You've probably read media stories about the results of opinion polls based on relatively small samples, for instance, "A survey of 1000 adults has shown that 35% of Canadians believe this country should not invest in any more nuclear power plants." It's quite remarkable that the statisticians in the survey company can select just 1000 people to be representative of the country's entire population. These organizations use surveys to answer such questions as

- How many people will accept our credit card with certain new features?
- How many Canadians who vote for our political party support the legalization of marijuana?

Statistics was successful in addressing these questions during the 20th century and will continue to excel in these areas during the 21st century, as shown in Figure 1.2. Now let's look at what's new in this century.

21st-Century Statistics

Today we continue to use statistics the way we did in the previous century, but with two major differences. First, much of the analysis is performed in real time, the moment the data become available; and second, the amounts of data available to us are much larger than ever before.

Real-Time Analysis of Data

According to IBM, "The biggest leaps forward in the next several decades—in business, science, and society at large—will come from insights gleaned through perpetual, real-time analysis of data. . . . The new science of analytics must be core to every leader's thinking."[3]

One example of what IBM refers to as "real-time analysis of data" is the way companies look at sales data in order to analyze their market. In the 20th century, these companies collected sales data at the end of each month and compiled them into reports for each region of the global market. Then they held quarterly and annual sales and marketing meetings at which regional directors shared their sales information with one another in order to identify patterns and trends. But by the time this was done, the results were often out of date. Today, companies record sales data in a database right when the product is sold, whether at the cash register in a retail store or when a salesperson signs a multimillion-dollar deal for industrial equipment. Those data are incorporated into a statistical analysis of global market trends that is immediately accessible to directors and executives throughout the company. In short, companies are now able to apply statistics in real time so that their analysis is completely up to date.

Analyzing Vast Amounts of Data

Corporate executives are keen to find useful value in the massive amounts of data now available to them. Even small companies can afford large databases and the statistical-analysis software that comes with them. So for this 21st-century revolution we've coined the term "data analytics" in order to focus on how all that data can be analyzed. And it's Statistics that provides a major methodology to tackle the problem. Moreover, Statistics is no longer being left to the statisticians; rather, it has become an increasingly important part of management decision making at all levels. Everywhere you look, statistics are being used in corporate planning, and this is why a solid grounding in Statistics is important for all managers.

Here are three examples of the results of analyzing really vast databases:

- Facebook gets more advertising revenue as a result of its members' visiting the site more frequently and actively contributing to their pages. The popular social network therefore tracked its members' behaviour using statistical analysis of its huge database—and found that the best predictor of whether members would contribute to the site was knowing that their friends had contributed. As a result of this analysis, Facebook started informing its members of what their friends had been saying.
- Some airlines routinely overbook flights because not all passengers show up. This is a delicate balancing act. The airlines don't want to lose revenue by flying with empty seats, but on the other hand they don't want to annoy passengers who are turned away and have to compensate them financially. If the airlines could improve their estimates of "no-shows," they'd be able to fine-tune how much overbooking they can do. *On average* they know the percentage of no-shows, but what about each individual flight, with its particular mix of passengers? Which passengers are the type who don't show? Statistical analysis allows airlines to match everything they know about each passenger with the number of times that passenger has been a no-show in the past. As a result of statistical analysis, one airline found that the passengers most likely to show up are those who order vegetarian meals. Now airlines take into account how many vegetarians they have on board when figuring out how much to overbook a particular flight.

[3] IBM. (2010). Building a smarter planet: 2 in a series: On a smarter planet, answers are hidden in the data. Retrieved from http://www.ibm.com/smarterplanet/global/files/us__en_us__intelligence__Data_visualization_4_6.pdf

- Closer to home, what can the Canadian winter teach retailers? They already know that if a storm results in a power outage, people will need batteries and flashlights. But statisticians have also found a correlation between storm warnings and sales of Pop-Tarts—a quick and easy snack you can eat even when the power is out. Now some retailers watch the weather forecast when deciding how much of that product to stock.

The Cutting Edge

In the three cases above, we knew the questions we were asking:

- How can we predict whether members will contribute to Facebook?
- How can an airline predict no-shows?
- Which products sell more during winter storms?

But the real challenge comes when a corporate executive does *not* have a specific question in mind, and instead asks management: "How can we improve our way of doing business by making use of our vast database of information and perhaps linking to other publicly available databases?" These more open-ended questions challenge us to think outside the box and apply statistical thinking in unusual ways.

Here's an example of how Google uses its own enormous database, along with a database from the European Union, to do language translation. If you ask Google to translate a document, say, from Spanish to Hungarian, it doesn't look each word up in a dictionary, in part because a single word in one language has many alternatives in another language. Instead, Google compares each phrase with phrases that appear in professionally translated European Union documents. The Google processor uses statistics to assess the probability of various possible translations of your phrase in its context, and then chooses the most likely one. And Google doesn't use statistics merely for language translation—statistics are at the core of its business. It continuously updates its analysis that ranks search results, taking into account evolving patterns in the various links people click on. Moreover, Google web crawlers select sites to "crawl" based on statistical analysis that chooses the sites most likely to have changed since they were last crawled.

We can gain competitive advantage in the 21st century by thinking outside the box and applying the full range of statistical analysis at our disposal to the vast databases that organizations are adding to every minute of every day.

LO1 1.2 How Is Statistics Used in Management?

Economic value has moved from goods to services and to data and the statistical algorithms used to analyse them.

—BASED ON *IT'S A SMART WORLD: A SPECIAL REPORT ON SMART SYSTEMS*

Statistical analysis is used to manage most public and private sector organizations, in just those areas that are popular with students in business schools: accounting, finance, marketing, and human resource planning.

Accounting

When a company's accounts are audited, the auditor often doesn't have the time to go through every item—for example, invoices. Instead, a "statistical audit" is conducted in which a representative sample of invoices is audited. The auditor then uses a statistical analysis of this sample to make valid conclusions about all the invoices to a required degree of accuracy. Chapters 11 to 17 are devoted to this topic, known as "statistical inference" since we are inferring a conclusion about all invoices from only a small sample of them.

"I keep saying that the sexy job in the next 10 years will be statisticians," says Hal Varian, chief economist at Google. "And I'm not kidding." "For Today's Graduate, Just One Word: Statistics."

—Steve Lohr,
The New York Times,
August 5, 2009

Finance

A major element in financial planning is managing risk. If you can measure something, you can manage it, and Statistics provides many ways of measuring risk. When an investor is choosing among alternative investments, he or she needs measures of their riskiness as well as their expected return on investment. These are statistical measures that we'll deal with in this book.

Marketing

Marketing, particularly retail marketing, is largely based on statistical analysis of consumer purchasing patterns. Most of Part 3 of this book is about the concept of regression, meaning how one variable relates to others, which is used to figure out how spending on a product depends on age group, income level, gender, postal code, and many other factors. This enables marketers to design promotional campaigns focused on the appropriate target audience.

Human Resource Planning

Any large organization today has a certain mix of employees at different levels in the management hierarchy. But what will that mix look like in 5 to 10 years' time? Will we have too many senior managers or not enough? The answer depends on statistical analysis of past patterns of promotion, recruitment, retirements, transfers, and resignations. Some of these, for example promotion and recruitment, are under the organization's control, but retirements and resignations are decisions made by employees for which we can calculate probabilities from past records. Part 2 of this book deals in detail with probabilities. Putting all this together enables us to calculate a statistical forecast of the number of employees at different levels of the management pyramid in the future.

1.3 How Can I Learn Statistics?

This book can teach you Statistics, but teaching isn't the same as learning. The book does the teaching, but you need to be very proactive in doing the learning by putting into practice the concepts and methods the book teaches. That's why we've provided you with MyStatLab. It is essential to practise examples of each learning objective of each chapter on MyStatLab, which includes many tools to help you, like "Help me solve this."

A coach teaches a hockey player how to play, but the player really acquires those skills only by practice on the ice. You learn Statistics in the same way as a hockey player learns hockey. This book is the coach, and the end-of-chapter exercises and MyStatLab are the ice. Statistics is like most useful things in life: You must practise it to really learn it.

How Will This Book Help?

That is a fair question. Most likely, this book will not turn out to be what you expect. It emphasizes graphics and understanding rather than computation and formulas. Instead of learning how to plug numbers into formulas, you'll learn the process of model development and come to understand the limitations of both the data you analyze and the methods you use. Every chapter uses real data and real business scenarios so that you can see how to use data to make decisions.

Netflix offered a $1 million prize in a competition to improve the company's movie recommendation software, and statistics was the main tool used by the contestants.

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.

—Sloppy Stats Shame Science
The Economist, June 3, 2004

This book includes numerous examples of the application of statistics in Canadian management situations. Canada is a major player internationally, and so to Canadian managers, international statistics are just as important as Canadian statistics. Our principal trading partner is, of course, the United States, so U.S. data are also of primary concern. Therefore, this book includes both U.S. and international business situations and data in addition to Canadian ones. You may choose a career in a Canadian company or in a multinational or in the public or nonprofit sectors. In that sense, this book mirrors the work environment of a typical Canadian business.

Graphs and Tables

Close your eyes and open this book at random. Is there a graph or table on the page? Do it again, say, 10 times. You probably saw data displayed in many ways, even near the back of the book and in the exercises. Graphs and tables help you understand what the data are saying. So each story and data set and every new statistical technique will come with graphics to help you understand both the methods and the data.

Optional Sections and Chapters

Some sections and chapters of this book are marked with an asterisk (*). These are optional, in the sense that subsequent material doesn't depend on them directly. We hope you'll read them anyway, as you did this section.

Getting Started

It's only fair to warn you: You can't get there by just reading the summaries. This book is different. It's not about memorizing definitions and learning equations. It's deeper than that. And much more interesting. But . . .

You have to read the book!

MINI case studies

Applications of Statistics in Business

Write one page describing an application of statistics in one of the functional areas of business (marketing, accounting, finance, . . .). Since this is Chapter 1, you are not expected to know which statistical method is appropriate. Instead, you should clearly state (i) the business problem to be solved, (ii) the data you expect to need in order to solve it, and (iii) the type of result that you might get from an analysis of those data. You can base your answer on an actual application of statistics by a specific organization or you can make up your own example.

Ioan Panaite/Shutterstock

2

Data

LEARNING OBJECTIVES

This chapter will show you how to probe data in order to understand it better. After reading and studying this chapter, you should be able to:

1. Identify the context of your data
2. Distinguish different types of data

Amazon.com

Amazon.com opened for business in July 1995, billing itself even then as "Earth's Biggest Bookstore," with an unusual business plan: Executives didn't plan to turn a profit for four to five years. Although some shareholders complained when the dot-com bubble burst, Amazon continued its slow, steady growth, becoming profitable for the first time in 2002. Since then, Amazon has remained profitable and has continued to grow.

It operates separate websites internationally, including the Canadian site Amazon.ca, which coordinates shipment from a fulfillment centre in Mississauga, Ontario. One key to Amazon's success is proprietary software that continuously analyzes data on past sales. Other businesses also use Amazon's unique analytical software. For instance, Sears Canada's website is powered by Amazon Services Canada and uses Amazon's software to track shopping patterns and other data. The results are used to give suggestions to Sears customers based on frequently purchased items and to provide comparison shopping among alternative brands.

Amazon R&D is constantly monitoring and revising its software to best serve customers and maximize sales performance. To make changes to the website, it experiments by collecting data and analyzing what works best. As Ronny Kohavi, former director of Data Mining and Personalization, 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."[1]

[1] Based on Amazon.com 2005 annual report; www.homegoodsonline.ca; www.sears.ca/gp/home.html. Accessed January 5, 2009.

The decision makers at Amazon.com recently stated, "Many of the important decisions we make at Amazon.com can be made with data. There is a right answer or a wrong answer, a better answer or a worse answer, and math tells us which is which. These are our favorite kinds of decisions."[2] It's clear that data analysis, forecasting, and statistical inference are at the core of the decision-making tools of Amazon.com.

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

—Used by permission of Ronny Kohavi.

Many years ago, store owners in small towns knew their customers personally. If you walked into the hobby shop, the owner might tell you about a new bridge that had come in for your Lionel train set. The tailor knew your dad's size, and the hairdresser knew how your mom liked her hair to be styled. There are still some stores like that around today, but we're increasingly likely to shop at large stores, by phone, or on the internet. Even so, when you phone an 800 number to buy new running shoes, customer service representatives may call you by your first name or ask about the socks you bought six weeks ago. Or the company may send an email in October offering new head warmers for winter running. That this same company can identify who you are, where you live, and the items you bought online—all without your even being asked to supply this information—is standard fare these days. How did the telephone sales representative know all these things about you?

The answer is data. Collecting data on customers, transactions, and sales lets companies track inventory and know what their customers prefer. These data can help businesses predict what their customers may buy in the future so that they'll know how much of each item to stock. And in connection with the earlier example, the store can use the data and what it learns from the data to improve customer service, mimicking the kind of personal attention a shopper experienced 50 years ago.

Companies use data to make decisions about other aspects of their business as well. By studying the past behaviour of customers and predicting their responses, they hope to better serve their customers and to compete more effectively. This process of using data, especially **transactional data** (data collected for recording a company's transactions), to make other decisions and predictions is sometimes called *data mining* or *predictive analytics.* The more general term **business analytics** (or sometimes simply *analytics*) describes *any* use of statistical analysis to drive business decisions from data, whether the purpose is predictive or simply descriptive.

LO❶ 2.1 What *Are* Data?

We bet you thought you knew this instinctively. Think about it for a minute. What exactly *do* we mean by **data**? Do data even have to be numbers? The amount of your last purchase in dollars is numerical data, but some data record names or other labels. The names in Amazon.com's database are regarded as data, but they are not numerical.

[2] From Amazon.com Annual Report. Published by amazon, © 2005.

THE FIVE W'S:
WHO
WHAT
WHEN
WHERE
WHY

Sometimes, data can have values that look like numerical values but are just numerals serving as labels. This can be confusing. For example, the ASIN (Amazon Standard Item Number) of a book may have a numerical value, such as 978-0321426592, but it's really just another *name* for the book *Business Statistics*.

Data values, no matter what kind, are useless without an understanding of their 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, they add *How* to the list as well. The situation is similar for statisticians. Answering these types of questions can provide a **context** for data values. The answers to the first two questions are essential. If you can't answer *Who* and *What*, you don't have data, and you don't have any useful information.

Table 2.1 shows an example of some of the data Amazon might collect:

10675489	B0000010AA	10.99	Chris G.	905	Quebec	15.98
Samuel P.	Nova Scotia	10783489	12837593	N	B000068ZVQ	15783947
Ontario	Katherine H.	16.99	Alberta	N	11.99	N
B000002BK9	902	Monique D.	Y	819	B0000015Y6	403

Table 2.1 An example of data with no context. It's impossible to say anything about what these values might mean without knowing their context.

Try to guess what the data in Table 2.1 represent. Why is that hard? Because these data have no *context*. We can make the meaning clear if we add the context of *Who* and *What* and organize the values into a **data table** such as the one in Table 2.2.

Purchase Order Number	Name	Ship to Province	Price	Area Code	Gift?	ASIN
10675489	Katherine H.	Alberta	10.99	403	N	B0000015Y6
10783489	Samuel P.	Nova Scotia	16.99	902	Y	B000002BK9
12837593	Chris G.	Quebec	15.98	819	N	B000068ZVQ
15783947	Monique D.	Ontario	11.99	905	N	B000001OAA

Table 2.2 Example of a data table. The variable names are in the top row. Typically, the *Who* of the table are found in the leftmost column.

Now we can see that the data in Table 2.2 represent four purchase records relating to orders from Amazon. The column titles tell *What* has been recorded. The rows tell us *Who*. But be careful. Look at all the variables to see *Who* the variables are about. Even if people are involved, they may not be the *Who* of the data. For example, the *Who* here are the purchase orders (not the people who made the purchases) because each row refers to a different purchase order, not necessarily a different *person*. A common place to find the *Who* of the table is the leftmost column. The other W's might have to come from the company's database administrator.[3]

In general, a row of a data table corresponds to an individual **case** about *Whom* (or about which—if they're not people) we record some characteristics. These cases go by different names, depending on the situation. An individual who answers a survey is referred to as a **respondent**. A person on whom we experiment is a **subject** or (in an attempt to acknowledge the importance of their role in the experiment) **participant**, but a company, website, or other inanimate subject is

[3] In database management, this kind of information is called "metadata," or data about data.

often called an **experimental unit**. In a database, a row is called a **record**—in this example, a purchase record. Perhaps the most generic term is *case*. In Table 2.2, the cases are the individual purchase orders.

Sometimes people refer to data values as *observations*, without being clear about the *Who*. Make sure you know the *Who* of the data, or you may not know what the data say. Each *characteristic* recorded about each individual or case is called a **variable**. These are usually shown as the columns of a data table, and they should have a name that identifies *What* has been measured. If the number of cases (*Who*) is smaller than the number of characteristics (*What*), we may interchange rows and columns so that *Who* is shown in columns and *What* is shown in rows.

A general term for a data table like this is a **spreadsheet**, a name that comes from bookkeeping ledgers of financial information. The data were typically spread across facing pages of a bound ledger, the book used by an accountant for keeping records of expenditures and sources of income. For the accountant, the columns were the types of expenses and income, and the cases were transactions, typically invoices or receipts.

Although data tables and spreadsheets are great for relatively small data sets, they're cumbersome for the complex data sets that companies must maintain on a day-to-day basis. And so various other architectures are used to store data, the most common being a relational database. In a **relational database**, two or more separate data tables are linked so that information can be merged across them. Each data table is a *relation* because it's about a specific set of cases with information about each of these cases for all (or at least most) of the variables ("fields" in database terminology). A table of customers, along with demographic information on each, is an example of such a relation. A data table with information about a different collection of cases is a different relation. For example, a data table of all the items sold by the company, including information on price, inventory, and past history, is a relation as well (as shown in Table 2.3). Finally, the day-to-day

Customers

Customer Number	Name	City	Province	Postal Code	Customer Since	Gold Member
473859	Rahini, R.	Magog	QC	J1X SV8	2007	No
127389	Li, V.	Guelph	ON	N1K 2H9	2000	Yes
335682	Marstas, J.	Calgary	AB	T2E 089	2003	No

Items

Product ID	Name	Price	Currently in Stock
SC5662	Silver Cane	43.50	Yes
TH2839	Top Hat	29.99	No
RS3883	Red Sequinned Shoes	35.00	Yes
. . .			

Transactions

Transaction Number	Date	Customer Number	Product ID	Quantity	Shipping Method	Free Ship?
T23478923	9/15/17	473859	SC5662	1	UPS 2nd Day	N
T23478924	9/15/17	473859	TH2839	1	UPS 2nd Day	N
T63928934	10/22/17	335473	TH2839	3	UPS Ground	N
T72348299	12/22/17	127389	RS3883	1	FedEx Ovnt	Y

Table 2.3 A relational database shows all the relevant information for the three separate relations linked by customer and product numbers.